Linking Learning Strategies and Performance for Support Vector Machines

Los Alamos
National Laboratory

James Howse, Don Hush, and Clint Scovel
Modeling, Algorithms, and Informatics, CCS-3
Mail Stop B265
Los Alamos National Laboratory
Los Alamos, NM 87545

{jhowse,dhush,jcs}@lanl.gov
505-665-2722

LANL Technical Report: LA–UR–02-1933

Report Date: May 22, 2002

Abstract

We develop a formal representation of the technique introduced in (Shawe-Taylor, Bartlett, Williamson, & Anthony, 1998; Shawe-Taylor & Cristianini, 1998) for bounding the generalization error of support vector machines. As a consequence we provide a framework that can be utilized to link learning strategies to their performance bounds in such a way that the bounds are expressed in terms of the structural properties of the learning strategy (e.g. characterizations of the optimum classifier in terms of the structure of the finite sample optimization criterion and its value at optimum). We use this framework to provide performance bounds for a class of support vector machines that includes the soft margin learning strategies commonly used in practice. We also show how to eliminate the effects of the center and scale of the data in the learning theorem. We apply this framework to improve results obtained in (Shawe-Taylor & Cristianini, 1998) for the 2-norm soft margin learning strategy by exploiting a relationship between covering numbers of classes of linear functions and covering numbers of linear operators. This result is expressed in terms of the finite sample criterion value at optimum. Finally we show how this bound can be expressed in terms of the random process.

1 Introduction

Vapnik's support vector machines provide a remarkably powerful and effective framework for classification (Vapnik, 1998). A rigorous analysis of their generalization error was first achieved in (Shawe-Taylor *et al.*, 1998; Shawe-Taylor & Cristianini, 1998). However there is much yet to be done with this analysis. For example the bounds in (Shawe-Taylor & Cristianini, 1998) appear to be inadequate for large sample size (see the beginning of Section 8 for more details). In addition we would like bounds that incorporate the fact that the classifier is determined by solving a soft margin optimization problem. Finally, although their technique appears quite powerful it is not clear how it can be applied to other learning strategies.

In this paper we develop a formal representation of the technique in (Shawe-Taylor & Cristianini, 1998). This facilitates the analysis of other learning strategies which we refer to collectively as support vector machines because of their similarity with the soft margin support vector machines of (Vapnik, 1998). It also allows incorporation of the fact that the classifier is a solution of a specific learning strategy. In addition we have improved their results by utilizing a relationship described in (Williamson, Smola, & Schölkopf, 2002) between covering numbers of classes of linear functions and covering numbers of linear operators. In particular this improvement provides performance guarantees with much better large sample behavior.

This paper is organized as follows. In Section 2 we define the σ-norm soft margin learning strategy. In Section 3 we derive a theorem (Theorem 2) providing bounds on the probability that there exists a function whose minimum value over a m–sample is at least γ larger than its value on a nontrivial fraction of future samples. Theorem 2 is based on a result in (Shawe-Taylor *et al.*, 1998) and forms the foundation for their technique. In Section 5 we present a general learning theorem (Theorem 3) which is obtained by applying the extension of (Shawe-Taylor & Cristianini, 1998) described in Section 4 to Theorem 2. In Section 6 we introduce the notion of *observables* as quantities which are available to the practitioner. We then present our main result (Theorem 5) describing the performance of an arbitrary learning strategy in terms of covering numbers of subsets of affine functions, where these subsets are determined by the learning strategy and the value of the observables. In Section 7 we describe the important case where the learning strategy is the minimization of one of the observables. In Section 8 we formalize the relationship described in (Williamson *et al.*, 2002) between covering numbers of classes of linear functions and covering numbers of linear operators. In Section 9 we describe how the σ-norm soft margin learning strategy fits into this general framework. This enables us to provide performance guarantees for the 2-norm soft margin learning strategy that have much better large sample behavior than in (Shawe-Taylor & Cristianini, 1998). These guarantees have the added benefit that they are expressed in terms of the 2-norm optimization criterion value at optimum and consequently express (implicitly) the influence of the free parameters of the optimization criterion on performance. We conclude Section 9 with a program for determining performance guarantees for the more general σ-norm soft margin learning strategies. In Section 10 we describe the interaction between symmetries of the learning strategy and prior information on Z. In particular we show how to eliminate the effects of the center and scale of Z in the learning theorem. Finally, in Section 11 we discuss the importance of learning strategies consisting of minimizing an empirical mean over the m-sample. In particular we show that the optimal value of the m-sample 2-norm criterion is concentrated *below* the optimal value

of the mean 2-norm criterion in terms of properties of the random variable Z. Consequently our performance guarantees for the 2-norm learning strategy can be expressed in terms of the optimal value of the mean 2-norm criterion and other functions of the random variable Z.

2 σ-norm soft margin support vector machines

Let X be a Banach space with dual space X^*. Consider a random variable $Z = (X, Y)$ with $Y = \{1, -1\}$. For simplicity we do not emphasize the difference between a Z-valued random variable and the space Z and we will try to not let this cause confusion. Given an m-sample $z^m = \{z_i, i = 1, .., m\}$, we design a linear classifier $y = sign(\psi \cdot x + b)$ with $\psi \in X^*$ and

$$sign(r) = 1, \ r \geq 0$$

$$sign(r) = -1, \ r < 0$$

by minimizing the σ-norm soft margin optimization criterion

$$J(\psi, b, \xi_1, .., \xi_m) = \frac{1}{m}\left(|\psi|^2 + \frac{1}{\Delta^\sigma}\sum_{i=1}^{m}\xi_i^\sigma\right) \tag{1}$$

with respect to (ψ, b) and ξ_i, subject to the constraints

$$\xi_i \geq 1 - y_i(\psi \cdot x_i + b) \tag{2}$$

$$\xi_i \geq 0. \tag{3}$$

For $\sigma \geq 1$ this is a convex programming problem. When $\sigma = 1$ or 2 it is a quadratic convex programming problem. We optimize with respect to $\xi_1, .., \xi_m$ for fixed ψ, b to obtain the equivalent unconstrained optimization problem which we call the σ-norm soft margin optimization problem with criterion

$$J_{z^m}(\psi, b) = \frac{1}{m}\left(|\psi|^2 + \frac{1}{\Delta^\sigma}\sum_{i=1}^{m}d^\sigma(z_i, \psi, b)\right) \tag{4}$$

with

$$d(z, \psi, b) = \max(1 - y(\psi \cdot x + b), 0). \tag{5}$$

It is possible to define the slack variables differently. For example, for $\gamma > 0$, let

$$d_\gamma(z, \psi, b) = \max(\gamma - y(\psi \cdot x + b), 0). \tag{6}$$

and let

$$J_{z^m,\gamma,\Delta}(\psi, b) = \frac{1}{m}\left(|\psi|^2 + \frac{1}{\Delta^\sigma}\sum_{i=1}^{m}d_\gamma^\sigma(z_i, \psi, b)\right) \tag{7}$$

be the σ-norm soft margin criterion. One obtains

$$J_{z^m,\gamma,\Delta}(\psi,b) = \gamma^2 J_{z^m,1,\gamma^{1-\frac{2}{\sigma}}\Delta}\left(\frac{\psi}{\gamma},\frac{b}{\gamma}\right)$$

so that $(\psi_*,b_*) \in \arg\min_{(\psi,b)} J_{z^m,\gamma,\Delta}(\psi,b)$ if and only if $(\frac{\psi_*}{\gamma},\frac{b_*}{\gamma}) \in \arg\min_{(\psi,b)} J_{z^m,1,\gamma^{1-\frac{2}{\sigma}}\Delta}(\psi,b)$. Since the classifier determined by $(\frac{\psi_*}{\gamma},\frac{b_*}{\gamma})$ is the same as that determined by (ψ_*,b_*), the classifier determined by the σ-norm soft margin optimization problem with a value of γ other than 1 can be implemented by modifying the choice of Δ to $\gamma^{1-\frac{2}{\sigma}}\Delta$ in the $\gamma = 1$ optimization criterion. Consequently, we only consider the slacks defined in equation 5.

As in Vapnik's theory of empirical risk minimization we want to provide bounds for the generalization error of classifiers built by optimizing the σ-norm soft margin criterion utilizing the fact that they are optimizers. In this paper we show that the technique developed by (Shawe-Taylor *et al.*, 1998) and (Shawe-Taylor & Cristianini, 1998) is well suited to this task. In the next section we derive the separation theorem that forms the foundation for their technique.

3 The separation theorem

We begin with some preparation. For any real valued function class $\mathcal{F}$ and any $t > 0$, let the squashed class $\pi_t(\mathcal{F})$ be defined by composition $f \mapsto \pi_t \circ f$ with the squashing function

$$\pi_t(s) = 0, \quad s < 0$$
$$\pi_t(s) = s, \quad 0 \le s \le t$$
$$\pi_t(s) = t, \quad s > t.$$

For a pseudometric space $(\mathcal{M},d)$ with pseudometric d, the covering number $\mathcal{N}(\epsilon,\mathcal{M},d)$ is the smallest number of open balls of radius ϵ that cover $\mathcal{M}$. For a class of functions $\mathcal{F}$ on a space Z with an m-sample z^m we use the pseudometric

$$d_{z^m}(f,g) = \max_{z\in z^m} |f(z) - g(z)|$$

and denote by $\mathcal{N}(\epsilon,\mathcal{F},d_{z^m})$ its covering numbers. When the possible values of the m-sample are constrained to a subset $\Omega \subset Z$ we let

$$\mathcal{N}(\epsilon,\mathcal{F},m,\Omega) = \sup_{z^m : z_i\in\Omega, i=1,..,m} \mathcal{N}(\epsilon,\mathcal{F},d_{z^m}) \tag{8}$$

denote the the largest covering numbers obtainable for Ω constrained m-samples.

The fundamental theorem that we use is a less general version of Lemma 4.6 from (Shawe-Taylor *et al.*, 1998).

Theorem 1. *Let Z be a random variable. Let $\gamma > 0$ be fixed and let $\mathcal{F}$ denote a class of real valued functions on Z. Suppose the support of Z is contained in the subset $\Omega \subset Z$. Let z^m denote the first half and w^m the second half of $2m$ iid samples. Then*

$$\mathcal{P}_{Z^m W^m}\left(\exists f \in \mathcal{F} : \min_i f(z_i) \ge \gamma, |\{i : f(w_i) \le 0\}| > m\epsilon\right) < \mathcal{N}\left(\frac{\gamma}{2},\pi_\gamma(\mathcal{F}),2m,\Omega\right) 2^{-m\epsilon}.$$

Proof. First we observe that

$$\mathcal{P}_{Z^m W^m}\Big(\exists f \in \mathcal{F}, \min_i f(z_i) \geq \gamma, |\{i : f(w_i) \leq 0\}| > m\epsilon\Big) =$$

$$\mathcal{P}_{Z^m W^m}\Big(\exists f \in \pi_\gamma(\mathcal{F}), \min_i f(z_i) \geq \gamma, |\{i : f(w_i) \leq 0\}| > m\epsilon\Big).$$

Consider a minimal $\gamma/2$ cover $B_{z^m w^m}$ of $\pi_\gamma(\mathcal{F})$ in the pseudometric $d_{z^m w^m}$. That is for every $f \in \pi_\gamma(\mathcal{F})$ there exists an $\tilde{f} \in B_{z^m w^m}$ such that $|\tilde{f}(z) - f(z)| < \gamma/2$ for all $z \in z^m w^m$. Consequently for any $f \in \pi_\gamma(\mathcal{F})$ with $\min_i f(z_i) \geq \gamma$ and $|\{i : f(w_i) \leq 0\}| > m\epsilon$ there is an $\tilde{f} \in B_{z^m w^m}$ with $\min_i \tilde{f}(z_i) > \gamma/2$ and $|\{i : \tilde{f}(w_i) < \gamma/2\}| > m\epsilon$ so that

$$\mathcal{P}_{Z^m W^m}\Big(\exists f \in \pi_\gamma(\mathcal{F}), \min_i f(z_i) \geq \gamma, |\{i : f(w_i) \leq 0\}| > m\epsilon\Big)$$

$$\leq \mathcal{P}_{Z^m W^m}\Big(\exists f \in B_{z^m w^m}, \min_i f(z_i) > \gamma/2, |\{i : f(w_i) < \gamma/2\}| > m\epsilon\Big)$$

For fixed f, we observe that the event $\{z^m w^m : \min_i f(z_i) > \gamma/2, |\{i : f(w_i) < \gamma/2\}| > m\epsilon\}$ implies that the smallest $m\epsilon$ samples must be in the right hand sample w^m. Consequently by introducing the permutation symmetries on $z^m w^m$ we observe that a fraction of at most $2^{-m\epsilon}$ of the sequences obtained through the permutations can satisfy the condition. Therefore, for fixed f,

$$\mathcal{P}_{Z^m W^m}\Big(\min_i f(z_i) > \gamma/2, |\{i : f(w_i) < \gamma/2\}| > m\epsilon\Big) \leq 2^{-m\epsilon}.$$

Consequently,

$$\mathcal{P}_{Z^m W^m}\Big(\exists f \in B_{z^m w^m} : \min_i f(z_i) > \gamma/2, |\{i : f(w_i) < \gamma/2\}| > m\epsilon\Big) \leq E(|B_{z^m w^m}|)2^{-m\epsilon}$$

and noting that $|B_{z^m w^m}| = \mathcal{N}(\frac{\gamma}{2}, \pi_\gamma(\mathcal{F}), d_{z^m w^m}) \leq \mathcal{N}(\frac{\gamma}{2}, \pi_\gamma(\mathcal{F}), 2m, \Omega)$, the proof is finished. ♦

By setting $\delta = \mathcal{N}(\frac{\gamma}{2}, \pi_\gamma(\mathcal{F}), 2m, \Omega)2^{-m\epsilon}$ we can write this result as

$$\mathcal{P}_{Z^m W^m}\Big(\exists f \in \mathcal{F}, \min_i f(z_i) \geq \gamma, |\{i : f(w_i) \leq 0\}| > m\epsilon(m, \delta)\Big) < \delta$$

where

$$\epsilon(m, \delta) = \frac{\log \mathcal{N}(\frac{\gamma}{2}, \pi_\gamma(\mathcal{F}), 2m, \Omega) + \log \frac{1}{\delta}}{m}.$$

We now move from the double sample to bounds on probabilities. We do this through an adaption of the ghost sample lemma of (Vapnik, 1998)

Lemma 1. *Let Z be a random variable and let $\mathcal{F}$ denote a class of functions on Z. Consider $2m$ iid samples, the first half denoted z^m and the second half w^m. Let $P_t(f \leq 0)$ be the fraction of the sample points w^m with $f \leq 0$. Let $C_f(\epsilon)$ denote the constant event $C_f(\epsilon) = \{\mathcal{P}_Z(f \leq 0) \geq \epsilon\}$ and $A^2_f(\epsilon)$ the event in the second variable $A^2_f(\epsilon) = \{w^m : P_t(f \leq 0) \geq \epsilon\}$. Then for any family B^1_f of events in the first variable z^m,*

$$\mathcal{P}_{Z^m}\Big(\cup_f (B^1_f \cap C_f(\epsilon))\Big) \leq 2\mathcal{P}_{Z^m W^m}\Big(\cup_f (B^1_f \cap A^2_f(\epsilon - \frac{1}{m}))\Big)$$

Proof. The proof follows that of (Vapnik, 1998)(relevant to the trivial case $B_f^1 = Z^m$) on page 132 very closely. ♦

Applying Lemma 1 to the result of Theorem 1 with

$$B_f^1 = \{z^m; \min_i f(z_i) \geq \gamma\}$$

we obtain the separation theorem

Theorem 2. *Let Z be a random variable. Let $\gamma > 0$ be fixed and let $\mathcal{F}$ denote a class of functions on Z. Suppose the support of Z is contained in the subset $\Omega \subset Z$. Consider m iid samples z^m. Then*

$$\mathcal{P}_{Z^m}\left(\exists f \in \mathcal{F}, \min_i f(z_i) \geq \gamma, \mathcal{P}_Z(f \leq 0) > \epsilon\right) < 4\mathcal{N}\left(\frac{\gamma}{2}, \pi_\gamma(\mathcal{F}), 2m, \Omega\right) 2^{-m\epsilon}.$$

The result can be restated; given $0 < \delta < 1$,

$$\mathcal{P}_{Z^m}\left(\exists f \in \mathcal{F}, \min_i f(z_i) \geq \gamma, \mathcal{P}_Z(f \leq 0) > \epsilon(m, \delta)\right) < \delta$$

where

$$\epsilon(m, \delta) = \frac{2 + \log \mathcal{N}(\frac{1}{2}, \pi_\gamma(\mathcal{F}), 2m, \Omega) + \log \frac{1}{\delta}}{m}.$$

4 Extension to separability

To use Theorem 2 for learning, we extend the basic variables, following (Shawe-Taylor & Cristianini, 1998), so that the event $\min_i f(z_i) \geq \gamma$ in the result of Theorem 2 is satisfied almost always. In this section we describe this extension.

Let $Z = (X, Y)$ and let $\Omega \subset Z$ contain the support of Z. Let Ω^m denote the product space containing the support of the m-sample spaces Z^m under iid sampling. Let V be a Banach space with its dual V^* and consider the direct sum Banach space

$$\hat{X} = X \times V$$

with norm $|(x, v)|^2 = |x|^2 + |v|^2$. It follows(see e.g.(Megginson, 1991)) that $\hat{X}^* = X^* \times V^*$ acts through

$$(x^*, v^*) \cdot (x, v) = x^* \cdot x + v^* \cdot v.$$

For any Banach space K, let $B_R(K) = \{k : 0 \leq |k| \leq R\}$ denote the ball of radius R. Given an m-sample k^m chosen from K we let $\mathfrak{B}(K^*)$ denote the space K^* of functions on K equipped the pseudometric d_{k^m}. We let $\mathfrak{B}(K^*)_R = B_R(K^*)$ and $\mathfrak{B}(K^*)_{r,s} = \{k : r < |k^*| \leq s\}$ denote the relevant pseudometric subspaces of $\mathfrak{B}(K^*)$. Let $\mathcal{A}(X) = \mathfrak{B}(X^*) + \Re$ denote the pseudometric space of affine functions on X, with linear part in $\mathfrak{B}(X^*)$, with pseudometric d_{x^m} and let

$\mathcal{A}(\hat{X}) = \mathfrak{B}(\hat{X}^*) + \Re$ denote the pseudometric space of affine functions on $\hat{X}$, with linear part in $\mathfrak{B}(\hat{X}^*)$, with pseudometric $d_{\hat{x}^m}$. We use the shorthand notation

$$\mathcal{A} = \mathcal{A}(\hat{X}) = X^* \times V^* \times \Re$$

and

$$\mathcal{L} = \mathfrak{B}(\hat{X}^*) = X^* \times V^*.$$

In addition to the standard notation (x^*, v^*, b) for a point in $\mathcal{A} = X^* \times V^* \times \Re$, for convenience we often use the notation

$$(\psi, \phi, b)$$

with $\psi \in X^*$, $\phi \in V^*$, and $b \in \Re$ and

$$(\Psi, b)$$

where $\Psi = (\psi, \phi) \in X^* \times V^*$ and $b \in \Re$. We also use the same notation (ψ, b) for a point in $\mathcal{A}(X) = X^* \times \Re$ and the classifier

$$y = sign(\psi \cdot x + b).$$

Consider maps

$$\kappa : X \to V \tag{9}$$

and

$$\kappa^* : X \to V^* \tag{10}$$

We use them to define extensions

$$E_X : X \to \hat{X}$$

by

$$E_X(x) = (x, \kappa_x) \tag{11}$$

and

$$E_{z^m} : \mathcal{A}(X) \to \mathcal{A}(\hat{X}) = \mathcal{A}$$

by

$$(\psi, b) \mapsto (\hat{\psi}_{z^m}, b) \tag{12}$$

where

$$\hat{\psi}_{z^m} = \Big(\psi, \sum_{i=1}^{m} y_i \kappa^*_{x_i} d\big((x_i, y_i), \psi, b\big)\Big). \tag{13}$$

We use the notation E_{z^m} to emphasize the dependence of this extension on the m-sample z^m. Let

$$\hat{Z} = (\hat{X}, Y)$$

and define the induced extension

$$E_Z : Z \to \hat{Z}$$

by

$$(x, y) \mapsto (E_X(x), y). \tag{14}$$

For any class $\mathcal{G}$ of functions on X or $\hat{X}$ we extend to a class

$$\mathcal{G} \mapsto \overline{\mathcal{G}} \tag{15}$$

on $Z = (X, Y)$ or $\hat{Z} = (\hat{X}, Y)$ by the folding

$$\overline{g}(z) = yg(x) \tag{16}$$

or

$$\overline{g}(\hat{z}) = yg(\hat{x}) \tag{17}$$

These extensions accomplish the following for the classes of functions $\mathcal{F} = \overline{\mathcal{A}(X)}$ and $\mathcal{F} = \overline{E_{z^m}\mathcal{A}(X)} \subset \overline{\mathcal{A}}$.

Lemma 2. *Let X and V be Banach spaces and let*

$$\hat{X} = X \times V$$

with norm $|(x, v)|^2 = |x|^2 + |v|^2$ denote the direct product and let $\hat{X}^* = X^* \times V^*$ denote its dual. Let $Z = (X, Y)$ and $\hat{Z} = (\hat{X}, Y)$. Suppose that there exist maps

$$\kappa : X \to V$$

and

$$\kappa^* : X \to V^*$$

such that $\kappa^*_{x_1} \cdot \kappa_{x_2} = 1$ if $x_1 = x_2$ and 0 otherwise. Consider an m-sample z^m where the x coordinates have no duplicates and define extensions E_X (line 11),E_Z (line 14) and E_{z^m} (lines 12 and 13).

If $z = (x, y)$ where x is not in the support of the x coordinate of the m-sample z^m, then

$$\overline{E_{z^m}(\psi, b)}\big(E_Z(z)\big) = \overline{(\psi, b)}(z)$$

and for $i = 1, .., m$,

$$\overline{E_{z^m}(\psi, b)}\big(E_Z(z_i)\big) \geq 1.$$

Proof. To prove the first we calculate

$$E_{z^m}(\psi,b)(E_X(x)) = \hat{\psi}\cdot E_X(x)+b = \psi\cdot x + \Big(\sum_{i=1}^{m} y_i\kappa^*_{x_i} d\big((x_i,y_i),\psi,b\big)\Big)\cdot\kappa_x + b = \psi\cdot x + b$$

for x not in the support of $x_i, i = 1,..,m$. Consequently

$$E_{z^m}(\psi,b)(E_X(x)) = (\psi,b)(x)$$

and from the definitions 16 and 17 of the foldings

$$\overline{E_{z^m}(\psi,b)}(E_Z(z)) = y\big(E_{z^m}(\psi,b)(E_X(x))\big) = y((\psi,b)(x)) = \overline{(\psi,b)}(z).$$

To prove the second we note that on a sample point (x_i,y_i)

$$\overline{E_{z^m}(\psi,b)}(E_Z(z_i)) = y_i\big(E_{z^m}(\psi,b)(E_X(x_i)\big) = y_i(\hat{\psi}_{z^m}\cdot E_X(x_i)+b) = y_i\Big(\psi\cdot x_i + y_i d\big((x_i,y_i),\psi,b\big)+b\Big)$$

$$= y_i(\psi\cdot x_i+b) + d\big((x_i,y_i),\psi,b\big) \geq 1$$

and the proof of Lemma 2 is finished. ♦

5 Abstract learning

In this section we combine the results of the last two sections to obtain a general learning theorem. We now consider a learning algorithm in a general way. Later we will be more specific. We first define the *learning strategy* $\mathfrak{O} : Z^m \to\to \mathcal{A}(X)$ which is a set-valued mapping

$$z^m \mapsto \mathfrak{O}_{z^m} \subset \mathcal{A}(X)$$

for some z^m dependent family $\mathfrak{O}_{z^m}$ of subsets of $\mathcal{A}(X)$. We refer to the subsets $\mathfrak{O}_{z^m}$ as *optima*. We call a *selection* $\mathfrak{L} : Z^m \to \mathcal{A}(X)$ from $\mathfrak{O}$ a learning algorithm where for each z^m

$$\mathfrak{L}_{z^m} \in \mathfrak{O}_{z^m}.$$

We write

$$(\psi_*,b_*) = \mathfrak{L}_{z^m} \tag{18}$$

as a shorthand notation for the solution produced by the learning algorithm. Here we do not concern ourselves with the computational efficiency of evaluating the function $\mathfrak{L}$. A way to think about the difference between $\mathfrak{L}$ and $\mathfrak{O}$ is that if the learning algorithm is specified in terms of minimizing an objective function J_{z^m} which depends upon the m-sample z^m, then $\mathfrak{O}_{z^m}$ is the set of all minimizers of J_{z^m} and the classifier $\mathfrak{L}_{z^m}$ is that point in $\mathfrak{O}_{z^m}$ produced by a specific algorithm chosen to perform this minimization.

Let

$$\hat{\mathfrak{O}}_{z^m} = E_{z^m}\mathfrak{O}_{z^m} = \{(\Psi,b) = E_{z^m}(\psi,b) : (\psi,b)\in\mathfrak{O}_{z^m}\} \tag{19}$$

denote the image of $\mathfrak{O}_{z^m}$ in $\mathcal{A}$ under the extension E_{z^m}. Let

$$\hat{\mathfrak{O}}_{\Omega^m} = \cup_{z^m \in \Omega^m} \hat{\mathfrak{O}}_{z^m} = \{(\Psi, b) = E_{z^m}(\psi, b) : (\psi, b) \in \mathfrak{O}_{z^m} \text{ for some } z^m \in \Omega^m\}$$

denote all possible images of optima under E_{z^m} as z^m varies over Ω^m. Let

$$\mathfrak{O}_{\Omega^m} = \cup_{z^m \in \Omega^m} \mathfrak{O}_{z^m}$$

denote all possible optima as z^m varies over Ω^m.

By applying the extensions, Lemma 2 allows Theorem 2 to be applied to the class of functions $\mathcal{F} = \overline{\hat{\mathfrak{O}}_{\Omega^m}}$ to provide performance bounds for the learning strategy $\mathfrak{O}$.

Theorem 3. *Let X be a Banach space and let $0 < \delta < 1$ be fixed. Consider a random variable $Z = (X, Y)$ with support contained in $\Omega \subset Z$, where X has no point mass. Let V denote a Banach space and consider the direct sum Banach space*

$$\hat{X} = X \times V$$

with norm $|(x, v)|^2 = |x|^2 + |v|^2$ and let $\hat{X}^* = X^* \times V^*$ denote its dual space. Suppose that there exist maps

$$\kappa : X \to V$$

and

$$\kappa^* : X \to V^*$$

such that $\kappa^*_{x_1} \cdot \kappa_{x_2} = 1$ if $x_1 = x_2$ and 0 otherwise. Define extensions E_X (line 11),E_Z (line 14) and E_{z^m}(lines 12 and 13). Let $\hat{\Omega} = E_Z\Omega$. Let (ψ_*, b_*) denote the solution produced by a learning algorithm $\mathfrak{L}$ which is a selection from a learning strategy $\mathfrak{O}$. Let $e(\psi, b) = \mathcal{P}_Z(y \neq sign(\psi \cdot x + b))$ denote the generalization error of the classifier $sign(\psi \cdot x + b)$.

Then

$$\mathcal{P}_{Z^m}\Big(e(\psi_*, b_*) > \epsilon(m, \delta)\Big) < \delta$$

where

$$\epsilon(m, \delta) = \frac{2 + \log \mathcal{N}\left(\frac{1}{2}, \pi_1(\overline{\hat{\mathfrak{O}}_{\Omega^m}}), 2m, \hat{\Omega}\right) + \log \frac{1}{\delta}}{m}.$$

Proof. Since the random variable X has no point mass the points of x^m will be unique with probability one and x will not be in the support of the fixed m-sample x^m with probability one. Therefore with probability one we can apply Lemma 2. Since an error in the classifier (ψ, b) at z implies that $\overline{(\psi, b)}(z) \leq 0$,

$$e(\psi, b) \leq \mathcal{P}_Z\big(\overline{(\psi, b)}(z) \leq 0\big). \tag{20}$$

Also the first part of Lemma 2 along with the assumption that X has no point mass implies that

$$\mathcal{P}_Z\big(\overline{(\psi, b)}(z) \leq 0\big) = \mathcal{P}_Z\big(\overline{E_{z^m}(\psi, b)}(\hat{z}) \leq 0\big). \tag{21}$$

Consequently

$$\mathcal{P}_{Z^m}\Big(e(\psi_*,b_*)>\epsilon(m,\delta)\Big)\leq \mathcal{P}_{Z^m}\Big(\mathcal{P}_Z\big(\overline{E_{z^m}(\psi_*,b_*)}(\hat{z})\leq 0\big)>\epsilon(m,\delta)\Big)$$

$$\leq \mathcal{P}_{Z^m}\Big(\exists(\Psi,b)\in\hat{\mathfrak{O}}_{z^m} \text{ and } \mathcal{P}_Z\big(\overline{(\Psi,b)}(\hat{z})\leq 0\big)>\epsilon(m,\delta)\Big)$$

$$= \mathcal{P}_{Z^m}\Big(\exists\overline{(\Psi,b)}\in\overline{\hat{\mathfrak{O}}_{z^m}} \text{ and } \mathcal{P}_Z\big(\overline{(\Psi,b)}(\hat{z})\leq 0\big)>\epsilon(m,\delta)\Big).$$

The second conclusion of Lemma 2 allows us to utilize Theorem 2 with $\gamma=1$ applied the class of functions $\overline{\hat{\mathfrak{O}}_{\Omega^m}}$ on $\hat{Z}$ as follows;

$$\mathcal{P}_{Z^m}\Big(\exists\overline{(\Psi,b)}\in\overline{\hat{\mathfrak{O}}_{z^m}} \text{ and } \mathcal{P}_Z\big(\overline{(\Psi,b)}(\hat{z})\leq 0\big)>\epsilon(m,\delta)\Big)$$

$$= \mathcal{P}_{Z^m}\Big(\exists\overline{(\Psi,b)}\in\overline{\hat{\mathfrak{O}}_{z^m}}:\overline{(\Psi,b)}(\hat{z}_i)\geq 1, i=1,..,m, \quad \text{and } \mathcal{P}_Z\big(\overline{(\Psi,b)}(\hat{z})\leq 0\big)>\epsilon(m,\delta)\Big)$$

$$\leq \mathcal{P}_{Z^m}\Big(\exists\overline{(\Psi,b)}\in\overline{\hat{\mathfrak{O}}_{\Omega^m}}:\overline{(\Psi,b)}(\hat{z}_i)\geq 1, i=1,..,m, \quad \text{and } \mathcal{P}_Z\big(\overline{(\Psi,b)}(\hat{z})\leq 0\big)>\epsilon(m,\delta)\Big)<\delta$$

where

$$\epsilon(m,\delta)=\frac{2+\log\mathcal{N}\left(\frac{1}{2},\pi_1(\overline{\hat{\mathfrak{O}}_{\Omega^m}}),2m,\hat{\Omega}\right)+\log\frac{1}{\delta}}{m}.$$

and the proof is finished. ♦

6 The main theorem

To characterize the performance of a learning algorithm using Theorem 3 we must understand how the structural properties of the learning strategy affect the covering numbers of $\pi_1(\overline{\hat{\mathfrak{O}}_{\Omega^m}})$. To have practical utility we would prefer that this relationship be expressed in terms of quantities that we can compute during the learning process. To this end we refine Theorem 3 so that the performance guarantees depend on quantities available to the practitioner that we call *observables*.

Consider a partition $\mathfrak{A}_k, k=1,..,K$ over $\hat{\mathfrak{O}}_{\Omega^m}$. That is, let

$$\mathfrak{A}_k\subset\mathcal{A}, k=1,..,K$$

satisfy

$$\cup_k\mathfrak{A}_k\supset\hat{\mathfrak{O}}_{\Omega^m}.$$

and

$$\mathfrak{A}_k\cap\mathfrak{A}_j=\emptyset, \quad k\neq j.$$

This definition implies the existence of a function $I:\hat{\mathfrak{O}}_{\Omega^m}\to\{1,..,K\}$ which designates which subset each point lies in.

We prove

Corollary 1. *With the assumptions of Theorem 3, consider a partition $\mathfrak{A}_k, k = 1, .., K$ over $\hat{\mathfrak{O}}_{\Omega^m}$. Let $I : \hat{\mathfrak{O}}_{\Omega^m} \to \{1, .., K\}$ denote the function which designates which subset each point lies in. Define*

$$\mathfrak{A}^*_{\Omega^m,k} = \hat{\mathfrak{O}}_{\Omega^m} \cap \mathfrak{A}_k. \tag{22}$$

Then

$$\mathcal{P}_{Z^m}\Big(e(\psi_*, b_*) > \epsilon\big(m, \delta, I(E_{z^m}(\psi_*, b_*))\big)\Big) < \delta$$

where

$$\epsilon(m, \delta, k) = \frac{2 + \log \mathcal{N}\left(\frac{1}{2}, \pi_1(\overline{\mathfrak{A}^*_{\Omega^m,k}}), 2m, \hat{\Omega}\right) + \log \frac{K}{\delta}}{m}.$$

Proof. We apply Theorem 3 to the substrategy defined by $E^{-1}_{z^m}\mathfrak{A}_k \cap \hat{\mathfrak{O}}_{z^m}$ to obtain

$$\mathcal{P}_{Z^m}\Big(E_{z^m}(\psi_*, b_*) \in \mathfrak{A}_k \text{ and } e(\psi_*, b_*) > \epsilon(m, \delta)\Big) < \delta$$

where

$$\epsilon(m, \delta) = \frac{2 + \log \mathcal{N}\left(\frac{1}{2}, \pi_1(\overline{\mathfrak{A}^*_{\Omega^m,k}}), 2m, \hat{\Omega}\right) + \log \frac{1}{\delta}}{m}.$$

Since the partition covers $\hat{\mathfrak{O}}_{\Omega^m}$

$$\mathcal{P}_{Z^m}\Big(e(\psi_*, b_*) > \epsilon\big(m, \delta, I(E_{z^m}(\psi_*, b_*))\big)\Big) = \mathcal{P}_{Z^m}\Big(\exists k : E_{z^m}(\psi_*, b_*) \in \mathfrak{A}_k \text{ and } e(\psi_*, b_*) > \epsilon(m, \delta, k)\Big)$$

and a union bound finishes the proof. ♦

We now choose the partition of Corollary 1 in terms of the values of observables. The observables that we consider are special in that they need to be pullbacks of functions under the extension E_{z^m}. More specifically, let

$$J : \mathcal{A} \to \mathcal{O}$$

be a map to a space $\mathcal{O}$ of observable values. Normally this space will be some $\Re^d$ but that is not necessary. J determines a function

$$J_{z^m} : \mathcal{A}(X) \to \mathcal{O} \tag{23}$$

through the pullback of the extension E_{z^m};

$$J_{z^m}(\psi, b) = J(E_{z^m}(\psi, b)) = J(\hat{\psi}_{z^m}, b). \tag{24}$$

Recall from equation 18 that we denote the solution produced by a learning algorithm $\mathfrak{L}$ by $(\psi_*, b_*) = \mathfrak{L}_{z^m}$. Let

$$J^*_{z^m} = J_{z^m}(\psi_*, b_*)$$

denote the value of the observable J_{z^m} at the solution provided by $\mathfrak{L}$.

We now specialize to $\mathcal{O} = \Re^+$. Although following analysis goes through in general, specializing to real valued observables makes the notation clearer. We now choose a specific partition in terms of the values of J. In particular, define

$$\mathcal{A}_{r,s} = \{(\Psi, b) \in \mathcal{A} : r < J(\Psi, b) \leq s\} \tag{25}$$

where

$$\mathcal{A}_{0^-,s} = \{(\Psi, b) \in \mathcal{A} : 0 \leq J(\Psi, b) \leq s\}. \tag{26}$$

to be the pullbacks under J of intervals in $\Re^+$. Then let

$$\mathcal{A}^*_{r,s} = \hat{\mathfrak{O}}_{\Omega^m} \cap \mathcal{A}_{r,s} \tag{27}$$

denote the intersection of these sets with $\hat{\mathfrak{O}}_{\Omega^m}$. That is $\mathcal{A}^*_{r,s}$ is the set of points $(\Psi, b) \in \mathcal{A}$ such that $(\Psi, b) = E_{z^m}(\psi, b)$ of some optima (ψ, b), for some m-sample $z^m \in \Omega^m$ and $J_{z^m}(\psi, b) = J(\hat{\psi}_{z^m}, b)$ lies between the appropriate interval. Notice that for convenience we have dropped the notational dependence on Ω, but that this dependence should be remembered.

We wish to choose a finite partition in terms of the value of $J^*_{z^m}$. So that such a partition will not include any sets with an infinite range of J^* values it is necessary that the range of J^* values be bounded. Without much loss in generality we suppose that such a bound

$$0 \leq J^*_{z^m} \leq M(J, \Omega), \quad \forall z^m \in \Omega^m \tag{28}$$

exists where the M depends on the learning algorithm, the observable J and the support Ω. In terms of $\mathcal{A}^*$ this bound implies that $\mathcal{A}^*_{r,s} = \emptyset$ for $r > M(J, \Omega)$.

Given the assumption 28 there are many ways to partition this interval inducing a partition of over $\hat{\mathfrak{O}}_{\Omega^m}$. We proceed as in (Shawe-Taylor & Cristianini, 1998). Lay down a partial arithmetic sequence

$$\jmath_k = \beta\alpha^{k-1}, k = 1, .., K$$

such that β is small, $\alpha > 1$ and $\jmath_K \geq M(J, \Omega)$. Denote

$$\jmath_0 = 0^-.$$

We can solve for

$$K = 1 + \left\lceil \frac{\log \frac{M(J,\Omega)}{\beta}}{\log \alpha} \right\rceil \leq 2 + \frac{\log \frac{M(J,\Omega)}{\beta}}{\log \alpha}.$$

Define

$$\mathfrak{A}_k = \mathcal{A}_{\jmath_{k-1},\jmath_k}, \quad k = 1, .., K \tag{29}$$

Applying this construction to Corollary 1 we obtain

Theorem 4. *With the assumptions of Theorem 3, consider the partition $\mathfrak{A}_k, k = 1, .., K$ defined above for some $\alpha > 1$ and $\beta > 0$. Let $I : \hat{\mathfrak{O}}_{\Omega^m} \to \{1, .., K\}$ denote the function which designates which subset each point lies in.*

Then

$$\mathcal{P}_{Z^m}\left(e(\psi_*, b_*) > \epsilon_1\big(m, \delta, I(E_{z^m}(\psi_*, b_*))\big)\right) < \delta$$

where

$$\epsilon_1(m, \delta, k) = \frac{2 + \log \mathcal{N}\left(\frac{1}{2}, \pi_1(\overline{\mathcal{A}^*_{\jmath_{k-1}, \jmath_k}}), 2m, \hat{\Omega}\right) + \log\left(2 + \frac{\log \frac{M(J,\Omega)}{\beta}}{\log \alpha}\right) + \log \frac{1}{\delta}}{m}.$$

This bound expresses the performance of the classifier (ψ_*, b_*) in terms of the observable value $J^*_{z^m}$ in the following way. The performance of the classifier (ψ_*, b_*) is expressed in terms of covering numbers of the subset $\mathcal{A}^*_{\jmath_{k-1}, \jmath_k}$ defined by $k = I(E_{z^m}(\psi_*, b_*))$. But in this case the definition of I is

$$I(E_{z^m}(\psi_*, b_*)) = k : J^*_{z^m} \in (\jmath_{k-1}, \jmath_k]$$

where the interval $(\jmath_{k-1}, \jmath_k]$ is $[0, \beta]$ when $k = 1$. Therefore this bound expresses the performance of the classifier in terms of two sequential values of J^* in the partial arithmetic sequence which contains $J^*_{z^m}$.

To make this expression depend more explicitly on $J^*_{z^m}$ we can proceed in the following way. When $J^*_{z^m} \in (\jmath_{k-1}, \jmath_k]$ and $k > 1$, $\jmath_{k-1} \geq \frac{J^*_{z^m}}{\alpha}$ and $\jmath_k < \alpha J^*_{z^m}$ and therefore

$$\mathcal{A}^*_{\jmath_{k-1}, \jmath_k} \subset \mathcal{A}^*_{\frac{J^*_{z^m}}{\alpha}, \alpha J^*_{z^m}}.$$

Said differently, when $k = I(E_{z^m}(\psi_*, b_*)) > 1$ then

$$\mathcal{A}^*_{\jmath_{k-1}, \jmath_k} \subset \mathcal{A}^*_{\frac{J^*_{z^m}}{\alpha}, \alpha J^*_{z^m}}.$$

If we split up the result of Theorem 4 into $I = 1$ and $I > 1$ and utilize the monotonicity of covering numbers we obtain

$$\mathcal{P}_{Z^m}\Big(\big(J^*_{z^m} > \beta, e(\psi_*, b_*) > \varepsilon(m, \delta, J^*_{z^m})\big) \quad \text{or} \quad \big(J^*_{z^m} \leq \beta, e(\psi_*, b_*) > \epsilon_*(m, \delta, \beta)\big)\Big) < \delta$$

where

$$\varepsilon(m, \delta, J^*_{z^m}) = \frac{2 + \log \mathcal{N}\left(\frac{1}{2}, \pi_1\left(\overline{\mathcal{A}^*_{\frac{J^*_{z^m}}{\alpha}, \alpha J^*_{z^m}}}\right), 2m, \hat{\Omega}\right) + \log\left(2 + \frac{\log \frac{M(J,\Omega)}{\beta}}{\log \alpha}\right) + \log \frac{1}{\delta}}{m}$$

and

$$\epsilon_*(m, \delta, \beta) = \frac{2 + \log \mathcal{N}\left(\frac{1}{2}, \pi_1\big(\overline{\mathcal{A}^*_{0^-, \beta}}\big), 2m, \hat{\Omega}\right) + \log\left(2 + \frac{\log \frac{M(J,\Omega)}{\beta}}{\log \alpha}\right) + \log \frac{1}{\delta}}{m}.$$

We apply the set theoretic identity $B_1 \cap B_2 \subset (A \cap B_1) \cup (A^c \cap B_2)$ to obtain

Theorem 5. *Let X be a Banach space and let $0 < \delta < 1$, $\alpha > 1$, and $\beta > 0$ be fixed. Consider a random variable $Z = (X, Y)$ with support contained in $\Omega \subset Z$, where X has no point mass. Let V denote a Banach space and consider the direct sum Banach space*

$$\hat{X} = X \times V$$

with norm $|(x, v)|^2 = |x|^2 + |v|^2$ and let $\hat{X}^* = X^* \times V^*$ denote its dual space. Suppose that there exist maps

$$\kappa : X \to V$$

and

$$\kappa^* : X \to V^*$$

such that $\kappa^*_{x_1} \cdot \kappa_{x_2} = 1$ if $x_1 = x_2$ and 0 otherwise. Define extensions E_X (line 11),E_Z (line 14) and E_{z^m}(lines 12 and 13). Let $\hat{\Omega} = E_Z\Omega$. Let (ψ_*, b_*) denote the solution produced by a learning algorithm $\mathfrak{L}$ which is a selection from a learning strategy $\mathfrak{O}$. Let a function $J : \mathcal{A} \to \Re^+$ define a a real valued observable through the pullback 24. Let the value of this observable at the solution be denoted $J^*_{z^m}$. Suppose there exists an $M(J, \Omega)$ such that $0 \le J^*_{z^m} \le M(J, \Omega), \quad \forall z^m \in \Omega^m$. Consider the sets defined in 27.

Then

$$\mathcal{P}_{Z^m}\left(e(\psi_*, b_*) > \max\left(\varepsilon(m, \delta, J^*_{z^m}), \epsilon_*(m, \delta, \beta)\right)\right) < \delta$$

where

$$\varepsilon(m, \delta, J^*_{z^m}) = \frac{2 + \log \mathcal{N}\left(\frac{1}{2}, \pi_1\left(\overline{\mathcal{A}^*_{\frac{J^*_{z^m}}{\alpha}, \alpha J^*_{z^m}}}\right), 2m, \hat{\Omega}\right) + \log\left(2 + \frac{\log \frac{M(J,\Omega)}{\beta}}{\log \alpha}\right) + \log \frac{1}{\delta}}{m}$$

and

$$\epsilon_*(m, \delta, \beta) = \frac{2 + \log \mathcal{N}\left(\frac{1}{2}, \pi_1\left(\overline{\mathcal{A}^*_{0^-, \beta}}\right), 2m, \hat{\Omega}\right) + \log\left(2 + \frac{\log \frac{M(J,\Omega)}{\beta}}{\log \alpha}\right) + \log \frac{1}{\delta}}{m}.$$

Theorem 5 has some free parameters available. It would also be nice to eliminate the $\epsilon_*(m, \delta, \beta)$ in the max inside the probability statement and have only $\varepsilon(m, \delta, J^*_{z^m})$. We will show for the 2-norm soft margin problem that we can determine a β that makes $\epsilon_*(m, \delta, \beta)$ extremely small so that we can effectively ignore it inside the statement while accounting for the price one pays in $\varepsilon(m, \delta, J^*_{z^m})$. In this case the price is negligible.

7 Learning through minimization

Theorem 5 only depends on the learning strategy $\mathfrak{O}$ and not on learning algorithm $\mathfrak{L}$ except through the condition that $\mathfrak{L}$ be a selection from $\mathfrak{O}$. Up until this point we have not said very

much about the choice of learning strategy $\mathfrak{O}$ or the choice of function J which determines observables J_{z^m}. We consider for the moment performance and ignore the important issue of computation. Let us first ignore the $\epsilon_*(m,\delta,\beta)$ term in the bound of Theorem 5. Let us consider for the moment trying to reduce the size $\mathcal{A}^*_{\frac{J^*_{z^m}}{\alpha},\alpha J^*_{z^m}}$ by defining the learning strategy which outputs the same solution (ψ_0, b_0) for every possible m-sample;

$$\mathfrak{O}_{z^m} = \mathfrak{L}_{z^m} = (\psi_0, b_0).$$

We do not think this is a good idea but want to see why not in the bounds. Theorem 3 expresses its performance in terms of the covering numbers of

$$\hat{\mathfrak{O}}_{\Omega^m} = \cup_{z^m \in \Omega^m} E_{z^m}(\psi_0, b_0)$$

but it is easy to see that despite the fact that we have one candidate classifier (ψ_0, b_0) the size of $\hat{\mathfrak{O}}_{\Omega^m}$ can be very large. Indeed we see from this example that we desire the strategy to be chosen so that the union of the images

$$z^m \mapsto \mathfrak{O}_{z^m} \mapsto E_{z^m}\mathfrak{O}_{z^m}$$

is small for the pseudometric determined by every other possible m-sample. This is only the beginning of an investigation into the question of what constitutes good learning strategy/observable pairs which we hope to continue in the future.

From the theorem we see that smaller $\mathcal{A}^*_{\frac{J^*_{z^m}}{\alpha},\alpha J^*_{z^m}}$ is better. Consider observables J_{z^m} where the difference between covering numbers of $\mathcal{A}^*_{r,s}$ and $\mathcal{A}^*_{0^-,s}$ is very small so we can use the latter in the result of Theorem 5. This is not an unreasonable condition. Indeed it is very similar to the fact that in large dimensional spaces the covering numbers of the unit sphere is not much less than the covering numbers of the whole unit ball. Since the covering numbers of $\mathcal{A}^*_{0^-,s}$ are monotonically decreasing as $s \downarrow 0$ and the performance bounds are now in terms of the covering numbers of $\mathcal{A}^*_{0^-,\alpha J_{z^m}(\psi,b)}$ we see that smaller J_{z^m} improves performance with the conclusion that we should consider letting our learning strategy be the minimization of J_{z^m}. In particular we choose a function J, pullback to the observable J_{z^m} and define

$$\mathfrak{O}_{z^m} = \arg \min_{(\psi,b)\in\mathcal{A}(X)} J_{z_m}(\psi, b)$$

to be the set of minima of J_{z_m}. We now refer to the observable J_{z^m} as an *optimization criterion*. Now the relevant sets are

$$\mathcal{A}^*_{r,s} = \{E_{w^m}(\psi,b) : (\psi,b) \ \text{ minimizes } \ J_{w^m}, J^*_{w^m} = J_{w^m}(\psi,b) \in (r,s], w^m \in \Omega^m\} \tag{30}$$

with $r = \frac{J^*_{z^m}}{\alpha}$ and $s = \alpha J^*_{z^m}$. For fixed z^m this is the union of the minima for J_{w^m} over the other m samples w^m with minimum criterion value close to $J^*_{z^m}$.

We can now ask how to choose the optimization criterion J_{z^m}. Consider a modified optimization criterion

$$\mathcal{J}_{z^m}(\psi,b) = \varepsilon(m,\delta,J_{z^m}(\psi,b)) = \frac{2 + \log \mathcal{N}\left(\frac{1}{2}, \pi_1(\overline{\mathcal{A}^*_{0^-,\alpha J_{z^m}(\psi,b)}}), 2m, \hat{\Omega}\right) + \log\left(2 + \frac{\log \frac{M(J,\Omega)}{\beta}}{\log \alpha}\right) + \log \frac{1}{\delta}}{m}.$$

Since this function has the same optima as J_{z^m} we get an almost perfect match; Let (ψ_*, b_*) denote a minimizer for $\mathcal{J}_{z^m}$, then

$$\mathcal{P}_{Z^m}\Big(e(\psi_*, b_*) > \mathcal{J}^*_{z^m}\Big) < \delta.$$

Consequently we can say that J is a good choice for the process Z if with high probability $\mathcal{J}^*_{z^m}$ is small.

The observation on line 30 facilitates the estimation of the covering numbers in the result of Theorem 5 when we use the learning strategy of minimizing the pullback(24) J_{z^m} of a function $J : \mathcal{A} \to \Re^+$ under the extension E_{z^m}. Indeed, Theorem 5 expresses this performance in terms of the covering numbers of $\pi_1\Big(\overline{\mathcal{A}^*_{\frac{J^*_{z^m}}{\alpha}, \alpha J^*_{z^m}}}\Big)$ and $\pi_1(\overline{\mathcal{A}^*_{0^-,\beta}})$ and the sets $\mathcal{A}^*_{\frac{J^*_{z^m}}{\alpha}, \alpha J^*_{z^m}}$ and $\mathcal{A}^*_{0^-,\beta}$ are the images under the extensions E_{z^m} of sets of optimal solution vectors for other m-sample values such that the optimal criterion value is within range specified by $J^*_{z^m}$. Consequently, it is here where we impose our knowledge about the optimal solution vectors specifically as a function of assumptions about the process Z, the function J, and the extensions E_{z^m} and E_X.

7.1 Consequences of minimization

With this in mind let us address the general problem of bounding the covering numbers

$$\mathcal{N}\left(\epsilon, \pi_1(\overline{\mathcal{A}^*_{r,s}}), 2m, \hat{\Omega}\right)$$

We first ignore the possible beneficial effects of considering the squashing function π_1 and bound these covering numbers by

$$\mathcal{N}\left(\epsilon, \pi_1(\overline{\mathcal{A}^*_{r,s}}), 2m, \hat{\Omega}\right) \leq \mathcal{N}\left(\epsilon, \overline{\mathcal{A}^*_{r,s}}, 2m, \hat{\Omega}\right).$$

Since in general

$$|\overline{f}(\hat{z}) - \overline{g}(\hat{z})| = |yf(\hat{x}) - yg(\hat{x})| = |f(\hat{x}) - g(\hat{x})|$$

if we suppose that

$$\hat{\Omega} \subset \big((\Omega_1, \Omega_2), Y\big)$$

with $\Omega_1 \subset X$ and $\Omega_2 \subset V$ then we can further bound these covering numbers by

$$\mathcal{N}\left(\epsilon, \overline{\mathcal{A}^*_{r,s}}, 2m, \hat{\Omega}\right) \leq \mathcal{N}\left(\epsilon, \mathcal{A}^*_{r,s}, 2m, (\Omega_1, \Omega_2)\right).$$

We do know something about these sets $\mathcal{A}^*_{r,s}$. For example, we know that

$$\mathcal{A}^*_{r,s} = \emptyset, \quad r > M(J, \Omega).$$

However we can say a little more

Lemma 3. *Suppose that J is monotonic in V^* with respect to κ^* in the sense that if $\phi_1 = \sum_{i=1}^k a_i^1 \kappa_{x_i}^* \in V^*$ and $\phi_2 = \sum_{i=1}^k a_i^2 \kappa_{x_i}^* \in V^*$ can be represented on a common set of unique points $x_i, i = 1, .., k$ so that $|a_i^1| \geq |a_i^2|$ for all i with strict inequality for at least one i, then $J(\psi, \phi_1, b) > J(\psi, \phi_2, b)$. Suppose (ψ_*, b_*) is a minimizer of J_{z^m}. If the sample data is all from one class y^*, choose $(\psi_*, b_*) = (0, y^*)$. Suppose also that $\Omega_1 \subset B_R(X)$. Then*

$$|b_*| \leq 1 + |\psi_*| R.$$

Proof. Consider first the case where the sample contains at least one point from each class. Suppose to the contrary of the lemma that $b_* > 1 + |\psi_*| R$. The argument for $b_* < -(1 + |\psi_*| R)$ is the same. Then for $y = 1$, $y(\psi_* \cdot x + b_*) = \psi_* \cdot x + b_* > 1$ so that $d((x, y), \psi_*, b_*) = \max(1 - y(\psi_* \cdot x + b_*), 0) = 0$. On the other hand for $y = -1$,

$$d((x, y), \psi_*, b_*) = 1 - y(\psi_* \cdot x + b_*) = 1 + \psi_* \cdot x + b_* > 2.$$

Since these inequalities are strict we can decrease b_* while not changing $d\big((x, y), \psi_*, b_*\big) = 0$ for sample data with $y = 1$ and strictly decreasing $d((x, y), \psi_*, b_*) = 1 + \psi_* \cdot x + b_*$ for sample data with $y = -1$. Since J is monotonic in V^* with respect to κ^*, this contradicts the optimality of $J_{z^m}(\psi, b) = \frac{1}{m}(|\psi|^2 + \frac{1}{\Delta^2} \sum_{i=1}^m d^2(z_i, \psi, b))$ at (ψ_*, b_*).

When the sample is all from one class y^*, it is easy to see that that $|b_*| = |y^*| \leq 1 + |\psi_*| R$ and the proof is finished. ♦

Although Lemma 3 provides a bound for b_* it is in terms of ψ_* which is not very useful in deriving bounds in terms of $J_{z^m}^*$. One further assumption alleviates this problem.

Lemma 4. *With the assumptions of Lemma 3, we suppose further that J is dominated from below in X^* in that there exists constants $c > 0$ and $\varrho > 0$ such that*

$$c|\psi|^\varrho \leq J(\psi, \phi, b)$$

for any (ψ, ϕ, b). Then

$$|b_*| \leq 1 + \left(\frac{J_{z^m}^*}{c}\right)^{\frac{1}{\varrho}} R.$$

Proof. The proof follows directly by applying the bound on $|b_*|$ deduced from Lemma 3 and then bounding ψ_* by $|\psi_*| \leq (\frac{J_{z^m}^*}{c})^{\frac{1}{\varrho}}$. ♦

We now want to utilize the result of Lemma 4. We use the following

Lemma 5. *Let Z be a topological space. Let $\mathcal{F} = \mathcal{F}_1 + \mathcal{F}_2$ be the direct sum of two classes of real valued functions on Z. Let z^m denote an m-sample from Z. Then for any $0 \leq \mu \leq 1$,*

$$\mathcal{N}(\epsilon, \mathcal{F}, d_{z^m}) \leq \mathcal{N}(\mu\epsilon, \mathcal{F}_1, d_{z^m}) \mathcal{N}((1 - \mu)\epsilon, \mathcal{F}_2, d_{z^m}).$$

Proof. For $f = f_1 + f_2$ and $g = g_1 + g_2$

$$d_{z^m}(f,g) = \max_{z \in z^m} |f(z) - g(z)| = \max_{z \in z^m} |f_1(z) + f_2(z) - g_1(z) - g_2(z)|$$

$$\leq \max_{z \in z^m} |f_1(z) - g_1(z)| + \max_{z \in z^m} |f_2(z) - g_2(z)| = d_{z^m}(f_1, g_1) + d_{z^m}(f_2, g_2)$$

Since generally $a + b \leq \sup\left(\frac{a}{\mu}, \frac{b}{1-\mu}\right)$ for any $0 \leq \mu \leq 1$ we obtain

$$d_{z^m}(f,g) \leq \max\left(\frac{d_{z^m}(f_1,g_1)}{\mu}, \frac{d_{z^m}(f_2,g_2)}{1-\mu}\right)$$

and consequently the product of an $\mu\epsilon$ covering of $\mathcal{F}_1$ with an $(1-\mu)\epsilon$ covering of $\mathcal{F}_2$ is an ϵ covering of $\mathcal{F}$ and the proof is finished. ⧫

To apply this result it is convenient to assume that J factors through $\mathcal{L}$ in that

$$J : \mathcal{A} \to \Re^+$$

is determined by extending a function

$$J_{\mathcal{L}} : \mathcal{L} \to \Re^+$$

to $\mathcal{A}$ by making it independent of the constants. We define

$$\mathcal{L}_{r,s} = \{\Psi \in \mathcal{L} : r < J_{\mathcal{L}}(\Psi) \leq s\} \tag{31}$$

where

$$\mathcal{L}_{0^-,s} = \{\Psi \in \mathcal{L} : 0 \leq J_{\mathcal{L}}(\Psi) \leq s\}. \tag{32}$$

The next lemma follows directly from Lemma 5 and shows how to incorporate information we know about the constant variables and the $\mathcal{L}$ variables of optimal solutions into the covering bounds.

Lemma 6. *Suppose that J factors through $\mathcal{L}$ and that $\mathcal{A}^*_{r,s} \subset M_{r,s} \times \Omega_{r,s}$ for families of subsets $\Omega_{r,s} \subset \Re$ and $M_{r,s} \subset \mathcal{L}$. Then for any $0 \leq \mu \leq 1$*

$$\mathcal{N}\left(\epsilon, \mathcal{A}^*_{r,s}, 2m, (\Omega_1, \Omega_2)\right) \leq \mathcal{N}\left(\mu\epsilon, M_{r,s}, 2m, (\Omega_1, \Omega_2)\right) \mathcal{N}\left((1-\mu)\epsilon, \Omega_{r,s}, |\cdot|\right)$$

where the rightmost term is the covering numbers of the set of real numbers $\Omega_{r,s}$ with respect to the usual metric.

We can now prove

Theorem 6. *Suppose we minimize the pullback 24 of a J which factors through $\mathcal{L}$ which is monotonic in V^* with respect to κ^* and dominated from below in X^* with constants c and ϱ as described in Lemma 4. If the sample data is all from one class y^*, choose the particular solution $(\psi_*, b_*) = (0, y^*)$. Suppose also that $\Omega_1 \subset B_R(X)$. Then for any $0 \leq \mu \leq 1$*

$$\mathcal{N}\left(\epsilon, \mathcal{A}^*_{r,s}, 2m, (\Omega_1, \Omega_2)\right) \leq \mathcal{N}\left(\mu\epsilon, \mathcal{L}_{r,s}, 2m, (\Omega_1, \Omega_2)\right) \left(\frac{2 + 2(\frac{s}{c})^{\frac{1}{\varrho}} R}{(1-\mu)\epsilon} + 1\right)$$

Proof. The proof follows directly from Lemma 6 using $M_{r,s} = \mathcal{L}_{r,s}$ and Lemma 4 and the fact that

$$\mathcal{N}(\epsilon, |b| \leq B, |\cdot|) \leq \frac{2B}{\epsilon} + 1.$$

♦

Theorem 6 incorporated the bounds on the constants from Lemma 4 but used only the crudest bound $M_{r,s} = \mathcal{L}_{r,s}$. We hope to incorporate sharper estimates $M_{r,s}$ in the future.

8 Covering numbers of linear functions

According to Theorem 5, to analyze the performance of a learning algorithm we need to analyze covering numbers of classes of affine functions. Theorem 6 states under what conditions these may be bounded in terms of the covering numbers of classes of linear functions. Although the results of (Alon, Ben-David, Cesa-Bianchi, & Haussler, 1997) are very general they do not take advantage of the linear structure of linear function classes. Indeed, application of the theorem of (Alon *et al.*, 1997) in the work of (Shawe-Taylor & Cristianini, 1998) provides bounds with an accuracy term of the form

$$\epsilon = O\left(H_{z^m} \log\left(\frac{1}{H_{z^m}}\right) \log(32m)\right)$$

where H_{z^m} is an empirical mean. If for large sample size H_{z^m} is concentrated around its mean, then it is easy to see that the $\log(32m)$ term causes this bound to become bad for large m. A similar observation has been made in (Graepel, Herbrich, & Williamson, 2001). We remedy this situation by mapping(following (Williamson *et al.*, 2002)) from linear function classes to linear operators so that we can apply the theory of covering numbers of linear operators. The results for the 2-norm soft margin problem are presented in Section 9.1 but motivated the following treatment. We require some preparation.

For a pseudometric space $(\mathcal{M}, d)$ the covering number $\mathcal{N}(\epsilon, \mathcal{M}, d)$ is the smallest number of open balls of radius ϵ that cover $\mathcal{M}$. The entropy numbers

$$\epsilon_n = \inf\{\epsilon : \mathcal{N}(\epsilon, \mathcal{M}, d) \leq n\}$$

are essentially the size of the smallest n balls that can cover $\mathcal{M}$. The dyadic entropy numbers are defined to be

$$e_n = \epsilon_{2^n - 1}.$$

The entropy numbers are essentially the inverse of the covering numbers, as illustrated in the following lemma.

Lemma 7. *Suppose that $e_n < f(n)$ were f is a strictly decreasing function on the natural numbers. Let $\epsilon_\infty = \lim_{n \mapsto \infty} f(n)$. Extend to a strictly decreasing function, also called f on the non-negative reals. Then*

$$\mathcal{N}(\epsilon, \mathcal{M}, d) \leq 2^{f^{-1}(\epsilon)}$$

for $\epsilon \geq \epsilon_\infty$.

Proof. The assumption of the lemma implies that $\epsilon_{2^{n-1}} < f(n)$. Consequently, there is an $\epsilon \leq f(n)$ such that $\mathcal{N}(\epsilon, \mathcal{M}, d) \leq 2^{n-1}$. By the monotonicity of the covering numbers $\mathcal{N}(f(n), \mathcal{M}, d) \leq 2^{n-1}$. For an arbitrary $\epsilon \in \Omega$, $\epsilon = f(f^{-1}(\epsilon)) \geq f(\lceil f^{-1}(\epsilon) \rceil)$ so that

$$\mathcal{N}(\epsilon, \mathcal{M}, d) \leq \mathcal{N}(f(\lceil f^{-1}(\epsilon) \rceil), \mathcal{M}, d) \leq 2^{\lceil f^{-1}(\epsilon) \rceil - 1} \leq 2^{f^{-1}(\epsilon)}$$

and the proof is finished. ♦

Therefore if we obtain a strict bound of $f(n)$ on the dyadic entropy numbers this translates to a bound of $2^{f^{-1}(\epsilon)}$ on the covering numbers.

Consider a linear operator $S : H_1 \to H_2$ between normed linear spaces. The entropy numbers of the operator S are defined to be the entropy numbers of the image of the unit ball. That is

$$e_n(S) = e_n(SB_1(H_1), d_2) \tag{33}$$

where $B_1(H_1)$ is the unit ball in H_1 and d_2 is the norm of H_2. In a similar way we define

$$\mathcal{N}(\epsilon, S) = \mathcal{N}(\epsilon, SB_1(H_1), d_2). \tag{34}$$

Let

$$\mathcal{N}(\epsilon, \Lambda \subset L(H_1, H_2)) = \sup_{S \in \Lambda} \mathcal{N}(\epsilon, S). \tag{35}$$

denote the definition analogous to equation 8 for a subset Λ of $L(H_1, H_2)$ the space of bounded linear operators from H_1 to H_2.

To analyze entropy numbers for linear function classes we construct the relevant linear operator as follows.

Lemma 8. *Let K be a Banach space. Consider an m-sample $k^m = \{k_i, i = 1, .., m\}$ constrained to $\Omega \subset K$. Denote the constraint on the m-samples $k^m \in \Omega^m$. Let $\mathcal{K} : K^* \to l_\infty^m$ denote the linear operator*

$$\mathcal{K}k^* = (k^* \cdot k_1, k^* \cdot k_2, .., k^* \cdot k_m). \tag{36}$$

Also let Ω^m denote the subset of linear operators $\mathcal{K} \in \Omega^m$ induced by the constraint $k^m \in \Omega^m$. Then for any $R > 0$

$$e_n(\mathfrak{B}(K^*)_R, d_{k^m}) = Re_n(\mathcal{K})$$

$$\mathcal{N}\left(\epsilon, \mathfrak{B}(K^*)_R, d_{k^m}\right) = \mathcal{N}\left(\frac{\epsilon}{R}, \mathcal{K}\right)$$

and

$$\mathcal{N}\left(\epsilon, \mathfrak{B}(K^*)_R, m, \Omega\right) = \mathcal{N}\left(\frac{\epsilon}{R}, \Omega^m \subset L(K^*, l_\infty^m)\right) \tag{37}$$

Proof. Consider two points $\mathcal{K}k_1^*$ and $\mathcal{K}k_2^*$ in the image of the unit ball $B_1(K^*)$. Then

$$d_{l_\infty^m}(\mathcal{K}k_1^*, \mathcal{K}k_2^*) = \max_i |(\mathcal{K}k_1^*)_i - (\mathcal{K}k_2^*)_i| = \max_i |k_1^* \cdot k_i - k_2^* \cdot k_i|$$

but the right hand side is equal to

$$d_{k^m}(k_1^*, k_2^*)$$

when considering k_1^* and k_2^* in $\mathfrak{B}(K^*)_1$. Consequently

$$e_n(\mathfrak{B}(K^*)_1, d_{k^m}) = e_n(\mathcal{K})$$

$$\mathcal{N}(\epsilon, \mathfrak{B}(K^*)_1, d_{k^m}) = \mathcal{N}(\epsilon, \mathcal{K})$$

and

$$\mathcal{N}(\epsilon, \Omega^m \subset L(K^*, l_\infty^m)) = \mathcal{N}(\epsilon, \mathfrak{B}(K^*)_1, m, \Omega) \tag{38}$$

follows from the definitions 35 and 8. By the homogeneity of the pseudonorm d_{k^m} and the space of linear functions($\mathfrak{B}(K^*)_R = R\mathfrak{B}(K^*)_1$),

$$e_n(\mathfrak{B}(K^*)_R, d_{k^m}) = Re_n(\mathfrak{B}(K^*)_1, d_{k^m})$$

and the proof is finished. ♦

Therefore, a bound on the entropy numbers $e_n(\mathcal{K})$ is a bound on the entropy numbers $e_n(\mathfrak{B}(K^*)_1, d_{k^m})$ and Lemma 7 can then be used to turn this into a bound on the covering numbers $\mathcal{N}(\epsilon, \mathfrak{B}(K^*)_1, d_{k^m})$.

9 Application to support vector machines

We now describe how the above theorems can be used for the σ-norm soft margin problems with criterion 4 with $1 \leq \sigma < \infty$. Consider $L_p(X) = L_p(X, \mathcal{D}, \mu)$, the space of all real valued Borel measurable functions with respect to the σ-algebra(no connection between this σ and that defining the σ-norm soft margin problem) of all subsets of X whose p-th power is integrable with respect to the counting measure μ defined by $\mu(x) = 1$ for all $x \in X$. $L_p(X)$ is a Banach space with dual space $L_q(X)$ where $\frac{1}{p} + \frac{1}{q} = 1$ when $1 \leq p < \infty$. When $p = \infty$, $L_\infty(X)$ is the space of bounded functions on X with sup norm and $L_\infty^*(X) = L_{bv}(X)$ the space of functions of bounded variation on X.(see e.g. (Yosida, 1978)).

Consider the choice $V = L_\tau(X), 1 < \tau \leq \infty$. We let $\kappa_x = \Delta\boldsymbol{\kappa}(x)$ and $\kappa_x^* = \frac{1}{\Delta}\boldsymbol{\kappa}(x)$ where $\boldsymbol{\kappa}$ is the function which $= 1$ at x and zero elsewhere and $\Delta > 0$ is some fixed constant.

Consider first $\tau \neq \infty$. Then $V^* = L_\sigma(X)$ where $\frac{1}{\sigma} + \frac{1}{\tau} = 1$. Consider the function

$$J : X^* \times L_\sigma(X) \times \Re \to \Re^+$$

defined by

$$J(\psi, \phi, b) = \frac{|\psi|^2 + |\phi|^\sigma}{m}. \tag{39}$$

The assumption of X having no point mass implies that the $x_i, i = 1, .., m$ are unique with probability one. Consequently, with probability one the optimization criterion defined through the pullback 24 of J defined in 39 by the extension E_{z^m} is

$$J_{z^m}(\psi, b) = J(\hat{\psi}_{z^m}, b) = \frac{|\psi|^2 + \left|\sum_{i=1}^m y_i \kappa^*_{x_i} d\big((x_i, y_i), \psi, b\big)\right|^{\sigma}_{L_\sigma}}{m} =$$

$$\frac{|\psi|^2 + \frac{1}{\Delta^\sigma}\sum_{i=1}^m d^\sigma\big((x_i, y_i), \psi, b\big)}{m}$$

which is the σ-norm soft margin optimization criterion 4.

When $\tau = \infty$ consider

$$J(\psi, \phi, b) = \frac{|\psi|^2 + |\phi|_{bv}}{m}.$$

The corresponding optimization criterion defined through the pullback 24 is

$$J_{z^m}(\psi, b) = J(\hat{\psi}_{z^m}, b) = \frac{|\psi|^2 + \left|\sum_{i=1}^m y_i \kappa^*_{x_i} d\big((x_i, y_i), \psi, b\big)\right|_{bv}}{m}.$$

Since the second term is a function of finite support it is in $L_1(X)$. It is generally true that V is isometrically embedded in V^{**} for any Banach space V (Yosida, 1978) so that $L_1(X)$ is isometrically embedded in $L_{bv}(X)$. Consequently we can compute the $L_{bv}(X)$ norm of the second term using the $L_1(X)$ norm. Therefore, using the fact that the $x_i, i = 1, .., m$ are unique with probability one we obtain that with probability one

$$J_{z^m}(\psi, b) = \frac{|\psi|^2 + \left|\sum_{i=1}^m y_i \kappa^*_{x_i} d\big((x_i, y_i), \psi, b\big)\right|_{L_1}}{m} = \frac{|\psi|^2 + \frac{1}{\Delta}\sum_{i=1}^m d\big((x_i, y_i), \psi, b\big)}{m}$$

which is the 1-norm soft margin optimization criterion 4. Therefore we can generate all the σ-norm soft margin optimization criteria in this framework.

Usually we do not concern ourselves with measurability issues but we feel at the minimum the extension E_Z should be measurable. In Lemma 13 in Appendix Appendix A: we prove that it is for most of the σ-norm soft margin problems. In addition, we now show that these criteria satisfy all of the additional assumptions required in the theorems and lemmas of this section. To begin with

$$J^*_{z^m} \leq J_{z^m}(0, 0) \leq \frac{1}{\Delta^\sigma}$$

for the σ-norm soft margin optimization criterion so we can choose

$$M(J, \Omega) = \frac{1}{\Delta^\sigma}$$

in the assumption of Theorem 4. In addition all the σ-norm J are monotonic in $L_\sigma(X)$ with respect to κ^* and dominated from below in X^* and so we can apply both Lemmas 3 and 4 to obtain bounds on the constant term in terms of the optimal criterion value $J^*_{z^m}$ if $\Omega_1 \subset B_R(X)$ with probability one. J factors through $\mathcal{L}$ so we can combine the bounds on the constants with bounds obtained for the linear part $J_{\mathcal{L}}$ using Lemma 5.

9.1 Bounds for the 2-norm soft margin support vector machine

We first consider the general case when

$$J_{\mathcal{L}}(x^*, v^*) = |x^*|^2 + |v^*|^2.$$

Then $J_{\mathcal{L}}(x^*, v^*) = |(x^*, v^*)|^2$ and consequently $\mathcal{L}_{0^-,s} = \mathfrak{B}(\hat{X}^*)_{\sqrt{s}}$ so the covering numbers of $\mathcal{L}_{0^-,s}$ can be computed directly in terms of the linear operator

$$(\mathcal{X}, \mathcal{V}) : \hat{X}^* \to l_\infty^{2m}$$

on the direct product Banach space through Lemma 8. If we apply the inequality $\mathcal{L}_{r,s} \subset \mathcal{L}_{0^-,s}$ then we see that all the covering numbers can be bounded in this way. We now show this technique works extremely well when both X and V are Hilbert spaces and apply this method to the 2-norm soft margin problem. In this case $\hat{X}^*$ is also a Hilbert space and the structure of Hilbert space allows us to deduce results about the covering numbers from results about $2m$ dimensional Hilbert space.

Theorem 7. *Let X be a Hilbert space and let $\Delta > 0$, $0 < \delta < 1$, $\alpha > 1$, $0 \leq \mu \leq 1$ and a small positive number $0 < \mathfrak{e} < \frac{1}{8}$ be fixed. Consider a random variable $Z = (X, Y)$, where X has no point mass and has support inside $B_R(X)$. Let z^m denote an iid m-sample. Let (ψ_*, b_*) denote a solution to the 2-norm soft margin optimization problem (4) with $\sigma = 2$ with optimal criterion value $J^*_{z^m} = J_{z^m}(\psi_*, b_*)$. If the sample data is all from one class y^*, choose the particular solution $(\psi_*, b_*) = (0, y^*)$. Let $e(\psi, b) = \mathcal{P}_Z(y \neq sign(\psi \cdot x + b))$ denote the generalization error of the classifier $sign(\psi \cdot x + b)$. Then for some constant $1.86 < C \leq 102.89$,*

$$\mathcal{P}_{Z^m}\Big(e(\psi_*, b_*) > \max\left(\varepsilon(m, \delta, J^*_{z^m}), \varepsilon_*(m, \delta, \mathfrak{e})\right)\Big) < \delta$$

where

$$\begin{aligned}\varepsilon(m, \delta, s) =& \frac{4\alpha C^2(R^2 + \Delta^2)s}{\mu^2} \log\left(1 + \frac{\mu^2}{2\alpha C^2(R^2 + \Delta^2)s}\right) \\ &+ \frac{\log\left(5 + 4(\alpha m s)^{\frac{1}{2}} R\right) + \log\left(2 + \frac{\log \frac{8C^2(R^2+\Delta^2)}{\mu^2 \mathfrak{e}^2 \Delta^2}}{\log \alpha}\right) + \log \frac{1}{\delta} + \log \frac{1}{1-\mu} + 2}{m}\end{aligned} \tag{40}$$

and

$$\varepsilon_*(m, \delta, \mathfrak{e}) = \mathfrak{e} + \frac{\log\left(5 + \sqrt{\mathfrak{e}}\frac{\mu}{C}\sqrt{\frac{m}{(R^2+\Delta^2)}}R\right) + \log\left(2 + \frac{\log \frac{8C^2(R^2+\Delta^2)}{\mu^2 \mathfrak{e}^2 \Delta^2}}{\log \alpha}\right) + \log \frac{1}{\delta} + \log \frac{1}{1-\mu} + 2}{m}. \tag{41}$$

Proof. Let $H = X \times L_2(X)$. From the previous section we know that with probability one the 2-norm optimization criterion equation 4 with $\sigma = 2$ is the pullback

$$J_{z^m}(\psi, b) = J(\hat{\psi}_{z^m}, b)$$

by the extension E_{z^m} of the function

$$J : X^* \times L_2^*(X) \times \Re \to \Re^+$$

defined by

$$J(\psi, \phi, b) = \frac{|\psi|^2 + |\phi|^2}{m}. \tag{42}$$

with $\kappa_x = \Delta\boldsymbol{\kappa}(x)$ and $\kappa_x^* = \frac{1}{\Delta}\boldsymbol{\kappa}(x)$. In addition, $0 \leq J_{z^m}^* \leq J_{z^m}(0,0) \leq 1$ so we can use

$$M(J, \Omega) = \frac{1}{\Delta^2}.$$

Consequently we can apply Theorem 5. To get meaningful bounds we need to bound the covering numbers

$$\mathcal{N}\left(\frac{1}{2}, \pi_1(\overline{\mathcal{A}_{r,s}^*}), 2m, \hat{\Omega}\right)$$

where $\Omega = ((\Omega_1, \Omega_2), Y)$ with

$$\Omega_1 = B_1(X)$$

and

$$\Omega_2 = B_\Delta(L_2(X)).$$

From the direct product structure in H

$$(\Omega_1, \Omega_2) \subset \Omega_H = B_{\sqrt{R^2+\Delta^2}}(H).$$

We go one step further than in Section 7 and use the sequence of inequalities

$$\mathcal{N}\left(\frac{1}{2}, \pi_1(\overline{\mathcal{A}_{r,s}^*}), 2m, \hat{\Omega}\right) \leq \mathcal{N}\left(\frac{1}{2}, \overline{\mathcal{A}_{r,s}^*}, 2m, \hat{\Omega}\right) \leq \mathcal{N}\left(\frac{1}{2}, \mathcal{A}_{r,s}^*, 2m, (\Omega_1, \Omega_2)\right) \leq \mathcal{N}\left(\frac{1}{2}, \mathcal{A}_{0^-,s}^*, 2m, \Omega_H\right).$$

Since J factors through $X^* \times L_2^*(X)$ and is monotonic in $L_2^*(X)$ with respect to $\kappa_x^* = \frac{1}{\Delta}\boldsymbol{\kappa}(x)$ and dominated from below in X^* with parameters $c = \frac{1}{m}$ and $\varrho = 2$ we can apply Lemma 6 to obtain that for any $0 \leq \mu \leq 1$

$$\mathcal{N}\left(\frac{1}{2}, \mathcal{A}_{0^-,s}^*, 2m, (\Omega_1, \Omega_2)\right) \leq \mathcal{N}\left(\frac{\mu}{2}, \mathcal{L}_{0^-,s}, 2m, (\Omega_1, \Omega_2)\right)\left(\frac{4 + 4(ms)^{\frac{1}{2}}R}{(1-\mu)} + 1\right). \tag{43}$$

Since

$$J(\psi, \phi, b) = \frac{|(\psi, \phi)|^2}{m} \tag{44}$$

it follows from the definition 32 that

$$\mathcal{L}_{0^-,s} = \mathfrak{B}(H^*)_{(ms)^{\frac{1}{2}}},$$

and so obtain

$$\mathcal{N}\left(\frac{\mu}{2}, \mathcal{L}_{0^-,s}, 2m, (\Omega_1, \Omega_2)\right) \leq \mathcal{N}\left(\frac{\mu}{2}, \mathfrak{B}(H^*)_{(ms)^{\frac{1}{2}}}, 2m, \Omega_H\right). \tag{45}$$

The following theorem is used to provide bounds on these covering numbers through the use of bounds on the covering numbers for linear operators. It appears that the utilization of the linear structure of the function class provides bounds which are superior to those obtainable by the more general results of (Alon *et al.*, 1997).

Theorem 8. *Let H be a Hilbert space. Consider an m-sample h^m constrained to $h_i \in \Omega = B_{R_1}(H), i = 1, .., m$. Let $R_2 > 0$ be fixed. There exists a constant $1.86 < C \leq 102.89$ such that when $\frac{C^2 R_1^2 R_2^2}{m\epsilon^2} \leq 1$*

$$\mathcal{N}(\epsilon, \mathfrak{B}(H^*)_{R_2}, m, \Omega) \leq 2^{\frac{C^2 R_1^2 R_2^2}{\epsilon^2} \log\left(1 + \frac{m\epsilon^2}{C^2 R_1^2 R_2^2}\right)}$$

Before we begin with the proof we note that the conjecture of (Williamson *et al.*, 2002) implies we can choose C as close to 1.86 as we wish.

Proof. We use a version of Maurey's Theorem, bounding $e_n(T)$ for any linear operator $T : l_2^m \to l_\infty^m$, obtained in (Williamson *et al.*, 2002) through an argument of Bernd Carl.

Theorem 9. *Let $T : l_2^m \to l_\infty^m$ be a linear operator. Then*

$$e_n(T) \leq C||T|| \left(\frac{\log\left(\frac{m}{n} + 1\right)}{n}\right)^{1/2}$$

with a constant $1.86 \leq C \leq 102.88$.

We will also use the following technical lemma

Lemma 9. *Suppose that $\eta \log(1 + \eta) = \xi \geq 1$ and $\eta \geq 0$. Then*

$$\frac{1}{\eta} \leq \frac{\log(1 + \xi)}{\xi}$$

Proof. Since the function $\eta \mapsto \eta \log(1 + \eta)$ is increasing for $\eta \geq 0$ and $\eta \log(1 + \eta) = 1$ when $\eta = 1$, $\xi \geq 1$ implies that $\eta \geq 1$. Consequently $\xi = \eta \log(1 + \eta) \geq \eta \log(1 + 1) = \eta$. Therefore, $\xi = \eta \log(1 + \eta) \leq \eta \log(1 + \xi)$ and the proof of Lemma 9 is finished. ♦

We now begin proof of Theorem 8. Consider the operator $\mathcal{H} : H \to l_\infty^m$ defined in equation 36. We need only consider nontrivial $\mathcal{H}$ because in that case the theorem in proven. Select any linear subspace $H_m \subset H$ of dimension m which contains the orthogonal complement to the kernel of $\mathcal{H}$. It is clear that $\mathcal{H}H = \mathcal{H}H_m$. Since this decomposition is orthogonal the length of $h + h_0$ for $h \in H_m$ and $h_0 \in H_m^\perp$ is minimal when $h_0 = 0$. Consequently in addition, $\mathcal{H}B_1(H) = \mathcal{H}B_1(H_m)$. Choose an orthonormal basis for H_m so that this basis determines the isometry $H_m = l_2^m$. From Theorem 9 we know that there is a constant $1.86 < C \leq 102.89$ such that $e_n(\mathcal{H}) < C||\mathcal{H}|_{H_m}|| \left(\frac{\log\left(\frac{m}{n}+1\right)}{n}\right)^{1/2}$ (the slight change in the range of C from above

is to generate a strict inequality in the bound). It is clear that $||\mathcal{H}|_{H_m}|| \leq ||\mathcal{H}||$ and since $h_i \in B_{R_1}(H), i = 1, .., m,$ we have $||\mathcal{H}|| \leq R_1$. Consequently,

$$e_n(\mathcal{H}) < CR_1 \left(\frac{\log\left(\frac{m}{n}+1\right)}{n} \right)^{1/2}.$$

From Lemma 8,

$$e_n(\mathfrak{B}(H^*)_{R_2}, d_{h^m}) = R_2 e_n(\mathcal{H}) < CR_1R_2 \left(\frac{\log\left(\frac{m}{n}+1\right)}{n} \right)^{1/2}.$$

Therefore we can apply Lemma 7 to

$$f(n) = CR_1R_2 \left(\frac{\log\left(\frac{m}{n}+1\right)}{n} \right)^{1/2}$$

with $\epsilon_\infty = 0$ and its natural extension to the non-negative reals. To do so we need to compute f^{-1}. To this end, let $\epsilon = f(n) = CR_1R_2\left(\frac{\log\left(\frac{m}{n}+1\right)}{n}\right)^{1/2}$. Then $\frac{\log\left(\frac{m}{n}+1\right)}{n} = \left(\frac{\epsilon}{CR_1R_2}\right)^2$ so that if we let $\eta = \frac{m}{n}$ and $\xi = m\left(\frac{\epsilon}{CR_1R_2}\right)^2$ then $\eta \log(1+\eta) = \xi$ and the assumption $\frac{C^2R_1^2R_2^2}{m\epsilon^2} \leq 1$ amounts to $\xi \geq 1$, so we can apply Lemma 9 to obtain that

$$\frac{1}{\eta} \leq \frac{\log(1+\xi)}{\xi}$$

and since $n = \frac{m}{\eta}$ we obtain

$$f^{-1}(\epsilon) \leq \frac{C^2R_1^2R_2^2}{\epsilon^2} \log\left(1 + m\left(\frac{\epsilon}{CR_1R_2}\right)^2\right).$$

We can now apply Lemma 7 to obtain

$$\mathcal{N}(\epsilon, \mathfrak{B}(H^*)_{R_2}, d_{h^m}) \leq 2^{\frac{C^2R_1^2R_2^2}{\epsilon^2}\log\left(1+m\left(\frac{\epsilon}{CR_1R_2}\right)^2\right)}$$

and the proof of Theorem 8 is finished. ♦

In succession we apply the bound 43 followed by inequality 45 followed by applying Theorem 8 with $\epsilon = \frac{\mu}{2}$, $R_1^2 = R^2 + \Delta^2$,$R_2 = (ms)^{\frac{1}{2}}$ and $m \mapsto 2m$. We obtain that for the C in Theorem 8, when $s \leq \frac{\mu^2}{2C^2(R^2+\Delta^2)}$

$$\mathcal{N}\left(\frac{1}{2}, \mathcal{A}^*_{0^-,s}, 2m, (\Omega_1, \Omega_2)\right) \leq \left(\frac{4+4(ms)^{\frac{1}{2}}R}{(1-\mu)} + 1\right) 2^{\frac{4C^2(R^2+\Delta^2)ms}{\mu^2}\log\left(1+\frac{\mu^2}{2C^2(R^2+\Delta^2)s}\right)}. \quad (46)$$

Consequently, by letting $s \mapsto \alpha s$ in the inequality 46, Theorem 5 tells us that

$$\mathcal{P}_{Z^m}\left(J^*_{z^m} \leq \frac{\mu^2}{2\alpha C^2(R^2+\Delta^2)} \text{ and } e(\psi_*, b_*) > \max\left(\epsilon(m, \delta, J^*_{z^m}), \epsilon_*(m, \delta, \beta)\right)\right) < \delta \quad (47)$$

where

$$\epsilon(m,\delta,s) = \frac{4\alpha C^2(R^2+\Delta^2)s}{\mu^2}\log\left(1+\frac{\mu^2}{2\alpha C^2(R^2+\Delta^2)s}\right) + \frac{\log\left(\frac{4+4(\alpha m s)^{\frac{1}{2}}R}{(1-\mu)}+1\right)+\log\left(2+\frac{\log\frac{1}{\beta\Delta^2}}{\log\alpha}\right)+\log\frac{1}{\delta}+2}{m} \tag{48}$$

where we suppress its formal dependence on β and

$$\epsilon_*(m,\delta,\beta) = \frac{4C^2(R^2+\Delta^2)\beta}{\mu^2}\log\left(1+\frac{\mu^2}{2C^2(R^2+\Delta^2)\beta}\right) + \frac{\log\left(\frac{4+4(m\beta)^{\frac{1}{2}}R}{(1-\mu)}+1\right)+\log\left(2+\frac{\log\frac{1}{\beta\Delta^2}}{\log\alpha}\right)+\log\frac{1}{\delta}+2}{m}. \tag{49}$$

The proof of Theorem 7 is essentially finished. We complete the proof in Appendix Appendix B: where we choose β so that $\epsilon_*(m,\delta,\beta)$ is so small that we don't mind it in $\max\left(\acute{\epsilon}(m,\delta,J^*_{z^m}),\epsilon_*(m,\delta,\beta)\right)$. ♦

We note that we can continue in this manner to choose α so that the second term

$$\frac{\log\left(5+\sqrt{\mathfrak{e}}\frac{\mu}{C}\sqrt{\frac{m}{(R^2+\Delta^2)}}R\right)+\log\left(2+\frac{\log\frac{8C^2(R^2+\Delta^2)}{\mu^2\mathfrak{e}^2\Delta^2}}{\log\alpha}\right)+\log\frac{1}{\delta}+\log\frac{1}{1-\mu}+2}{m}.$$

in the definition (41) of $\varepsilon_*(m,\delta,\mathfrak{e})$ is small. Having done so would guarantee that the term

$$\frac{\log\frac{1}{\delta}+\log\frac{1}{1-\mu}+2}{m}$$

in the definition (40) of $\varepsilon(m,\delta,s)$ is small. However we will not carry out these details here.

9.2 Bounds for σ-norm and other learning strategies

When $\sigma \neq 2$ we cannot proceed in such a straightforward manner as we did in Section 9.1. As discussed at the beginning of this section, the results of the previous sections apply. In particular E_Z is measurable and we can choose $M(J,\Omega) = \frac{1}{\Delta^\sigma}$. In addition all the σ-norm criteria are monotonic in $L_\sigma(X)$ with respect to κ^* and dominated from below in X^* and so we can apply both lemma 3 and 4 to obtain bounds on the constant term in terms of the optimal criterion value $J^*_{z^m}$ if $\Omega_1 \subset B_R(X)$. J factors through $\mathcal{L}$ so we can combine the bounds on the constants with bounds obtained for the linear part $J_{\mathcal{L}}$ using Lemma 5. Consequently, to provide specific bounds as we did in the 2-norm case what is left is to provide bounds on the covering numbers of $\mathcal{L}_{r,s}$.

In Appendix Appendix C: we describe a program for analyzing the covering numbers of $\mathcal{L}_{r,s}$ in terms of covering numbers of operators from $X^* \to l_\infty^{2m}$ and $V^* \to l_\infty^{2m}$. We do this for a large class of criteria containing the σ-norm soft margin criteria.

10 Symmetries in J and prior information about Z

All of the previous bounds were in terms of a set $\Omega \subset Z$ which contained the support of the random variable Z. Consequently, if we do not know Ω then we will not know the bounds. On the other hand if we know enough about the support that we can construct a set which contains the support then we can use this set in these bounds. We now consider a learning problem where we have some information about Z and we want to use that information to determine something about Δ. We show that if the function J has some symmetry properties that we can determine a dependence of Δ on Z so that we can remove the scale dependence of the bounds. We now make Δ a function of the random variable Z and formally indicate Δ's dependence on Z as $\Delta(Z)$ where now $\Delta : Z_{rv} \mapsto \Re^+$ is a real valued function on the space Z_{rv} of Z-valued random variables. It is important to note that Δ is not a real valued function on the base space (also noted Z) of the random variables. Most of the discussion below can be carried out in general, but to be specific we stick to the 2-norm soft margin problem.

$$J_{z^m,Z}(\psi, b) = \frac{1}{m}\left(|\psi|^2 + \frac{1}{\Delta^2(Z)}\sum_{i=1}^{m} d^2(z_i, \psi, b)\right) \tag{50}$$

with

$$d(z, \psi, b) = \max(1 - y(\psi \cdot x + b), 0). \tag{51}$$

Consider the affine transformation $X \mapsto r\mathcal{O}X + x_0$ where r is a scalar, $\mathcal{O}$ is orthogonal, and x_0 is a constant. We denote m copies of this transformation by $X^m \mapsto r\mathcal{O}X^m + x_0^m$. Let

$$\arg\min_{(\psi,b)} J_{z^m,Z}(\psi, b)$$

denote all solutions of the 2-norm soft margin optimization problem.

Lemma 10. *Suppose that*

$$\Delta\big((r\mathcal{O}X + x_0, Y)\big) = r\Delta\big((X, Y)\big).$$

If $(\Psi, B) \in \arg\min_{(\psi,b)} J_{z^m,Z}(\psi, b)$, *then*

$$\left(\frac{1}{r}\mathcal{O}\Psi, B - \frac{1}{r}\mathcal{O}\Psi \cdot x_0\right) \in \arg\min_{(\psi,b)} J_{(r\mathcal{O}x+x_0,y)^m,(r\mathcal{O}X+x_0,Y)}(\psi, b).$$

Proof. Since $\Delta\big((r\mathcal{O}X + x_0, Y)\big) = r\Delta\big((X, Y)\big)$

$$J_{(r\mathcal{O}x+x_0,y)^m,(r\mathcal{O}X+x_0,Y)}(\psi, b) = \frac{1}{m}\left(|\psi|^2 + \frac{1}{\Delta^2(r\mathcal{O}X + x_0, Y)}\sum_{i=1}^{m} d^2\big((r\mathcal{O}x_i + x_0, y_i), \psi, b\big)\right)$$

$$= \frac{1}{m}\left(|\psi|^2 + \frac{1}{r^2\Delta^2(Z)}\sum_{i=1}^{m} d^2\big((r\mathcal{O}x_i + x_0, y_i), \psi, b\big)\right)$$

but since $\psi \cdot (r\mathcal{O}x + x_0) + b = r\mathcal{O}^t\psi \cdot x + (\psi \cdot x_0 + b)$ and $< \mathcal{O}\psi, \mathcal{O}\psi >= |\psi|^2$ if we let $\acute{\psi} = r\mathcal{O}^t\psi$ this becomes

$$\frac{1}{mr^2}\left(|\acute{\psi}|^2 + \frac{1}{\Delta^2(Z)}\sum_{i=1}^{m} d^2\left((x_i, y_i),\ \acute{\psi},\ \frac{1}{r}\mathcal{O}\acute{\psi} \cdot x_0 + b\right)\right) = \frac{1}{r^2}J_{z^m,Z}(\acute{\psi}, \frac{1}{r}\mathcal{O}\acute{\psi} \cdot x_0 + b)$$

Since

$$\min_{(\psi,b)} J_{z^m,Z}(\psi, \frac{1}{r}\mathcal{O}\psi \cdot x_0 + b) = \min_{(\psi,b)} J_{z^m,Z}(\psi, b)$$

the result follows. ♦

Consequently, if the function Δ has the symmetry property required by Lemma 10 then the classifiers determined by the soft margin problem are equivariant to translating and normalizing the data. This important fact can remove the dependence of the performance on scale choice. Instead of stating a general theorem in this regards we choose a specific function and show its effect for the 2-norm soft margin support vector machine.

Define

$$\Delta(Z) = \Theta \inf_{x_0,r}\{r : supp(X) \subset B(x_0, r)\}$$

for some $\Theta > 0$ where $\inf_{x_0,r}\{r : supp(X) \subset B(x_0, r)\}$ is the radius of the smallest ball containing the support of X. Then it is clear that Δ satisfies the covariance condition $\Delta\big((r\mathcal{O}X + x_0, Y)\big) = r\Delta\big((X, Y)\big)$. In addition, if $R = \inf_{x_0,r}\{r : supp(X) \subset B(x_0, r)\}$ then $\Delta(Z) = \Theta R$ and we can translate $X \mapsto \acute{X} = \frac{1}{R}X - \frac{1}{R}x_0$ so that in the $\acute{X}$ variables

$$\Delta\big((\acute{X}, Y)\big) = \Theta$$

.

Let (ψ_*, b_*) be a optimizer of $J_{z^m,Z}$. From Lemma 10, $(\acute{\psi}, \acute{b}) = (R\psi_*, b_* + \psi_* \cdot x_0)$ is an optimizer of $J_{(\acute{x},y)^m,(\acute{X},Y)}$. Since $\acute{\psi} \cdot \acute{x} + \acute{b} = \psi \cdot x + b$

$$\mathcal{P}_Z(sign(\psi \cdot x + b) \neq y) = \mathcal{P}_{\acute{X},Y}(sign(\acute{\psi} \cdot \acute{x} + \acute{b}) \neq y)$$

so that we can apply Theorem 7 to obtain

Theorem 10. *Let X be a Hilbert space and let $\Theta > 0$, $0 < \delta < 1$, $\alpha > 1$, $0 \leq \mu \leq 1$ and a small positive number $0 < \mathfrak{e} < \frac{1}{8}$ be fixed. Consider a random variable $Z = (X, Y)$, where X has no point mass and define $R = \inf_{x_0,r}\{r : supp(X) \subset B(x_0, r)\}$ and $\Delta(Z) = \Theta R$ where $B(x_0, r)$ is the ball of radius r centered at the point $x_0 \in X$. Let z^m denote an iid m-sample. Let (ψ_*, b_*) denote a solution to the 2-norm soft margin optimization problem (4) with optimal criterion value $J^*_{z^m} = J_{z^m}(\psi_*, b_*)$. If the sample data is all from one class y^*, choose the particular solution $(\psi_*, b_*) = (0, y^*)$. Let $e(\psi, b) = \mathcal{P}_Z(y \neq sign(\psi \cdot x + b))$ denote the generalization error of the classifier $sign(\psi \cdot x + b)$. Consider the centered and scaled variable $\acute{x} = \frac{1}{R}(x - x_0)$ where x_0 denotes the center of the smallest ball which contains the support of X.*

Let $\acute{J}_{z^m} = \min_{(\psi,b)} \frac{1}{m}(|\psi|^2 + \frac{1}{\Theta^2}\sum_{i=1}^m d^2((\acute{x}_i, y_i), \psi, b))$ *denote the optimal value in the centered normalized coordinates.*

Then for some constant $1.86 < C \leq 102.89$,

$$\mathcal{P}_{Z^m}\Big(e(\psi_*, b_*) > \max\left(\varepsilon(m, \delta, \acute{J}_{z^m}), \varepsilon(m, \delta, \mathfrak{e})_*\right)\Big) < \delta$$

where

$$\begin{aligned}\varepsilon(m,\delta,s) =& \frac{4\alpha C^2(1+\Theta^2)s}{\mu^2}\log\left(1+\frac{\mu^2}{2\alpha C^2(1+\Theta^2)s}\right)\\ &+\frac{\log\left(5+4(\alpha m s)^{\frac{1}{2}}\right)+\log\left(2+\frac{\log\frac{8C^2(1+\Theta^2)}{\mu^2\mathfrak{e}^2\Theta^2}}{\log\alpha}\right)+\log\frac{1}{\delta}+\log\frac{1}{1-\mu}+2}{m}\end{aligned} \tag{52}$$

and

$$\varepsilon_*(m,\delta,\mathfrak{e}) = \mathfrak{e} + \frac{\log\left(5+\sqrt{\mathfrak{e}}\frac{\mu}{C}\sqrt{\frac{m}{(1+\Theta^2)}}\right)+\log\left(2+\frac{\log\frac{8C^2(1+\Theta^2)}{\mu^2\mathfrak{e}^2\Theta^2}}{\log\alpha}\right)+\log\frac{1}{\delta}+\log\frac{1}{1-\mu}+2}{m}. \tag{53}$$

11 Minimizing empirical means

An important property of the σ-norm soft margin criteria is that they are defined by empirical means. To be more precise, let

$$\mathfrak{p}_{z^m}$$

be the empirical measure associated with the m-sample z^m. We denote $E_{\mathfrak{p}_{z^m}}$ the process of taking the empirical mean. In this section E_Z denotes the expectation with respect to the random variable Z, as opposed to the rest of the paper where it is an extension. Then the σ-norm criterion can be written

$$J_{z^m}(\psi, b) = \frac{|\psi|^2}{m} + \frac{1}{\Delta^\sigma} E_{\mathfrak{p}_{z^m}}(d^\sigma(z, \psi, b)).$$

Since $E_{Z^m} E_{\mathfrak{p}_{z^m}} = E_Z$ we obtain

$$E_{Z^m}(J_{z^m}(\psi, b)) = \frac{|\psi|^2}{m} + \frac{1}{\Delta^\sigma} E_Z(d^\sigma(z, \psi, b)).$$

Define the function

$$J_Z(\psi, b) = E_{Z^m}(J_{z^m}(\psi, b)) = \frac{|\psi|^2}{m} + \frac{1}{\Delta^\sigma} E_Z(d^\sigma(z, \psi, b)). \tag{54}$$

Let (ψ_Z, b_Z) denote a minimizer of J_Z with minimum value J_Z^* and (ψ_{z^m}, b_{z^m}) denote a minimizer of J_{z^m} with minimum value $J_{z^m}^*$. Then

$$\begin{aligned} J_{z^m}^* - J_Z^* &= J_{z^m}(\psi_{z^m}, b_{z^m}) - J_Z(\psi_Z, b_Z) \leq J_{z^m}(\psi_Z, b_Z) - J_Z(\psi_Z, b_Z) \\ &= \frac{1}{\Delta^\sigma}\left(E_{\mathfrak{p}_{z^m}}(d^\sigma(z, \psi_Z, b_Z)) - E_Z(d^\sigma(z, \psi_Z, b_Z))\right) \end{aligned} \tag{55}$$

and

$$\begin{aligned} J_Z^* - J_{z^m}^* &= J_Z(\psi_Z, b_Z) - J_{z^m}(\psi_{z^m}, b_{z^m}) \leq J_Z(\psi_{z^m}, b_{z^m}) - J_{z^m}(\psi_{z^m}, b_{z^m}) \\ &= \frac{1}{\Delta^\sigma}\left(E_Z(d^\sigma(z, \psi_{z^m}, b_{z^m})) - E_{\mathfrak{p}_{z^m}}(d^\sigma(z, \psi_{z^m}, b_{z^m}))\right) \\ &\leq \frac{1}{\Delta^\sigma} \sup_{(\psi,b)\in\mathfrak{O}_\Omega} \left(E_Z(d^\sigma(z, \psi, b)) - E_{\mathfrak{p}_{z^m}}(d^\sigma(z, \psi, b))\right) \end{aligned} \tag{56}$$

Putting bounds 55 and 56 together implies that

$$|J_Z^* - J_{z^m}^*| \leq \frac{1}{\Delta^\sigma} \sup_{(\psi,b)\in\mathfrak{O}_\Omega^+} |E_Z(d^\sigma(z, \psi, b)) - E_{\mathfrak{p}_{z^m}}(d^\sigma(z, \psi, b))| \tag{57}$$

where $\mathfrak{O}_\Omega^+ = \mathfrak{O}_\Omega \cup (\psi_Z, b_Z)$ so we can bound

$$\mathcal{P}_{Z^m}(|J_Z^* - J_{z^m}^*| \geq \epsilon) \leq \mathcal{P}_{Z^m}\left(\sup_{(\psi,b)\in\mathfrak{O}_\Omega^+} |E_Z(d^\sigma(z, \psi, b)) - E_{\mathfrak{p}_{z^m}}(d^\sigma(z, \psi, b))| \geq \epsilon\Delta^\sigma\right).$$

The righthand side can be bounded by theorems on the uniform converge of empirical means in terms of the covering numbers of the set of functions $d^\sigma(z, \psi, b)$ determined as (ψ, b) varies over $\mathfrak{O}_\Omega^+$. See (Vapnik, 1998) for example. However, what is more important is that the one sided bound 55 can be used to bound

$$\mathcal{P}_{Z^m}(J_{z^m}^* \geq J_Z^* + \epsilon) \leq \mathcal{P}_{Z^m}\left(E_{\mathfrak{p}_{z^m}}(d^\sigma(z, \psi_Z, b_Z)) - E_Z(d^\sigma(z, \psi_Z, b_Z)) \geq \epsilon\Delta^\sigma\right)$$

which can be bounded in terms of the convergence of empirical means but this time the function $d^\sigma(z, \psi_Z, b_Z)$ is fixed because (ψ_Z, b_Z) is not a random variable. Therefore there will be no function class complexity term and if we can bound the size of (ψ_Z, b_Z) then we can use the Hoeffding bounds (see e.g. (McDiarmid, 1991)) and this probability should be very small. We can prove the following for the 2-norm strategy.

Theorem 11. *With the assumptions of Theorem 7, let $0 < \acute{\delta} < 1$ be fixed and let $J_\infty(\psi, b) = \frac{1}{\Delta^2} E_Z(d^2(z, \psi, b))$ be the mean infinite sample 2-norm criterion and let (ψ_∞, b_∞) one of its minima. Let μ_∞ denote the root of the second moment of the margin of the errors of (ψ_∞, b_∞). In addition, let the mean m-sample 2-norm criterion J_Z, defined as in 54, have a minimum value J_Z^*.*

Then for some constant $1.86 < C \leq 102.89$,

$$\mathcal{P}_{Z^m}\Big(e(\psi_*, b_*) > \max\left(\varepsilon(m, \delta, J_Z^* + \epsilon), \varepsilon(m, \delta, \mathfrak{e})_*\right)\Big) < \delta + \acute{\delta}$$

where

$$\varepsilon(m,\delta,s) = \frac{4\alpha C^2(R^2+\Delta^2)s}{\mu^2}\log\left(1+\frac{\mu^2}{2\alpha C^2(R^2+\Delta^2)s}\right) + \frac{\log\left(5+4(\alpha m s)^{\frac{1}{2}}R\right)+\log\left(2+\frac{\log\frac{8C^2(R^2+\Delta^2)}{\mu^2\mathfrak{e}^2\Delta^2}}{\log\alpha}\right)+\log\frac{1}{\delta}+\log\frac{1}{1-\mu}+2}{m}, \tag{58}$$

$$\varepsilon_*(m,\delta,\mathfrak{e}) = \mathfrak{e} + \frac{\log\left(5+\sqrt{\mathfrak{e}}\frac{\mu}{C}\sqrt{\frac{m}{(R^2+\Delta^2)}}R\right)+\log\left(2+\frac{\log\frac{8C^2(R^2+\Delta^2)}{\mu^2\mathfrak{e}^2\Delta^2}}{\log\alpha}\right)+\log\frac{1}{\delta}+\log\frac{1}{1-\mu}+2}{m}, \tag{59}$$

and

$$\epsilon = \epsilon(\acute{\delta}, m, R, \mu_\infty) = \frac{4\left(1+\frac{R}{\mu_\infty}\right)^2}{\Delta^2}\sqrt{\frac{\log\frac{1}{\acute{\delta}}}{m}}$$

Proof. As mentioned above, to apply the Hoeffding inequality we need to bound $d^2(z,\psi_Z,b_Z)$. We proceed first by bounding ψ_Z in terms of ψ_∞.

Lemma 11. *Consider the mean σ-norm soft margin criteria in* 54. *Let (ψ_{m_1},b_{m_1}) and (ψ_{m_2},b_{m_2}) denote minimizers of J_Z when $m=m_1$ and $m=m_2$ respectively. Suppose that $1\le m_1\le m_2\le\infty$. Then*

$$|\psi_{m_1}|\le|\psi_{m_2}|.$$

Proof. We suppose to the contrary that $|\psi_{m_1}|>|\psi_{m_2}|$ and write the abbreviation

$$J_Z(\psi,b)=\frac{|\psi|^2}{m}+D(\psi,b).$$

Then

$$\frac{|\psi_{m_2}|^2}{m_1}+D(\psi_{m_2},b_{m_2}) = \left(\frac{1}{m_1}-\frac{1}{m_2}\right)|\psi_{m_2}|^2+\frac{|\psi_{m_2}|^2}{m_2}+D(\psi_{m_2},b_{m_2})$$

$$< \left(\frac{1}{m_1}-\frac{1}{m_2}\right)|\psi_{m_1}|^2+\frac{|\psi_{m_2}|^2}{m_2}+D(\psi_{m_2},b_{m_2})$$

but since (ψ_{m_2},b_{m_2}) is optimal when $m=m_2$, $\frac{|\psi_{m_2}|^2}{m_2}+D(\psi_{m_2},b_{m_2})\le\frac{|\psi_{m_1}|^2}{m_2}+D(\psi_{m_1},b_{m_1})$ and we obtain

$$\left(\frac{1}{m_1}-\frac{1}{m_2}\right)|\psi_{m_1}|^2+\frac{|\psi_{m_2}|^2}{m_2}+D(\psi_{m_2},b_{m_2})\le\left(\frac{1}{m_1}-\frac{1}{m_2}\right)|\psi_{m_1}|^2+\frac{|\psi_{m_1}|^2}{m_2}+D(\psi_{m_1},b_{m_1})$$

$$=\frac{|\psi_{m_1}|^2}{m_1}+D(\psi_{m_1},b_{m_1})$$

which contradicts the optimality of (ψ_{m_1},b_{m_1}) and the proof is finished. ♦

If we let $m = m_1$ and denote as before $(\psi_Z, b_Z) = (\psi_m, b_m)$ and consider (ψ_∞, b_∞). Then Lemma 11 implies that

$$|\psi_Z| \leq |\psi_\infty|. \tag{60}$$

With a slight modification of the proof of Lemma 3 we conclude $|b_Z| \leq 1+|\psi_Z|R \leq 1+|\psi_\infty|R$ so that

$$(\psi_Z, b_Z)$$

is bounded in terms of $|\psi_\infty|$ uniformly in m. Consequently we desire to bound $|\psi_\infty|$.

For a fixed (ψ, b) let the map to margin be written

$$z \mapsto \rho(z) = \frac{1}{|\psi|} y(\psi \cdot x + b)$$

and let the induced probability measure on margin be denoted $\mathfrak{p}_\rho$.

Lemma 12. *Let (ψ_∞, b_∞) denote a minimizer of the mean infinite sample 2-norm soft margin criterion J_∞. Let $\mathfrak{p}_\rho$ denote the margin distribution measure for (ψ_∞, b_∞), and let*

$$\mu_{2,s} = \sqrt{\frac{\int_{-\infty}^{s} \rho^2 d\mathfrak{p}_\rho}{\int_{-\infty}^{s} 1 d\mathfrak{p}_\rho}}$$

be the root of the second moment of the margin less than s. Suppose that $\mathfrak{p}_\rho$ is absolutely continuous. Then

$$|\psi_\infty| \leq \frac{1}{\mu_{2,0}}.$$

Proof. Let $(\psi_\infty, b_\infty) = r(\Psi, B)$ with $|\Psi| = 1$ and $r = |\psi_\infty|$. We write

$$E_Z(d^2(z, \psi, b)) = \int_{-\infty}^{\frac{1}{r}} (1 - r\rho)^2 d\mathfrak{p}_\rho$$

where $d\mathfrak{p}$ denotes the integration process. Since (ψ_∞, b_∞) is a minimizer, r should be at a minimum and since $\mathfrak{p}_\rho$ is absolutely continuous, a critical point with respect to r is characterized by

$$\int_{-\infty}^{\frac{1}{r}} (1 - r\rho)\rho d\mathfrak{p}_\rho = 0.$$

except at the endpoint $r = 0$ in which case the lemma is proven. Solving for

$$r = \frac{\int_{-\infty}^{\frac{1}{r}} \rho d\mathfrak{p}_\rho}{\int_{-\infty}^{\frac{1}{r}} \rho^2 d\mathfrak{p}_\rho}. \tag{61}$$

and applying Schwarz's inequality

$$\int_{-\infty}^{\frac{1}{r}} \rho d\mathfrak{p}_\rho \leq \Big(\int_{-\infty}^{\frac{1}{r}} \rho^2 d\mathfrak{p}_\rho\Big)^{\frac{1}{2}} \Big(\int_{-\infty}^{\frac{1}{r}} 1 d\mathfrak{p}_\rho\Big)^{\frac{1}{2}}$$

implies that

$$r \leq \frac{1}{\mu_{2,\frac{1}{r}}}.$$

Since $\mu_{2,s}$ is monotonically increasing in s, $\mu_{2,\frac{1}{r}} \geq \mu_{2,0}$ and the proof Lemma 12 is finished. ♦

Noting that $\mu_\infty = \mu_{2,0}$, Lemma 12 proves that $|b_Z| \leq 1 + \frac{R}{\mu_\infty}$ so that

$$0 \leq d^2(z, \psi_Z, b_Z) \leq 4(1 + \frac{R}{\mu_\infty})^2$$

and the Hoeffding inequality (McDiarmid, 1991) can be applied to obtain

$$\mathcal{P}_{Z^m}(J^*_{z^m} - J^*_Z \geq \epsilon) \leq \mathcal{P}_{Z^m}\left(E_{\mathfrak{p}_{z^m}}(d^2(z, \psi_Z, b_Z)) - E_Z(d^2(z, \psi_Z, b_Z)) \geq \epsilon\Delta^2\right)$$

$$\leq e^{-m\frac{\epsilon^2\Delta^4}{16(1+\frac{R}{\mu_\infty})^4}}$$

and the proof of Theorem 11 follows easily. ♦

12 Acknowledgments

We would like to thank Bernd Carl for many helpful comments and thank Leonid Gurvits for the idea of using the covering numbers of linear operators for linear function classes.

References

Alon, N., Ben-David, S., Cesa-Bianchi, N., & Haussler, D. (1997). Scale-sensitive dimensions, uniform convergence, and learnability. *Journal of the ACM*, *44*(4), 615–631.

Carl, B. (1985). Inequalities of Bernstein-Jackson-type and the degree of compactness of operators in Banach spaces. *Ann. Inst. Fourier, Grenoble*, *35*(3), 79–118.

Graepel, T., Herbrich, R., & Williamson, R. C. (2001). From margin to sparsity. *Advances in Neural Information Processing*, *13*(?), in press.

McDiarmid, C. (1991). Concentration. In Habib, M., McDiarmid, C., Ramirez-Alfonsin, J., & Reed, B. (Eds.), *Probabilistic Methods for Algorithmic Discrete Mathematics*, pp. 195–248. Springer.

Megginson, R. E. (1991). *An Introduction to Banach Space Theory.* Springer-Verlag, New York.

Shawe-Taylor, J., Bartlett, P., Williamson, R. C., & Anthony, M. (1998). Structural Risk Minimization over Data-Dependent Hierarchies. *IEEE Transactions on Information Theory*, *44*(5), 1926–1940.

Shawe-Taylor, J., & Cristianini, N. (1998). Robust bounds on generalization from the margin distribution. *NeuroCOLT technical report, NC-TR-1998-020, ESPRIT NeuroCOLT2 Working Group, http://www.neurocolt.com.*

Vapnik, V. N. (1998). *Statistical Learning Theory.* John Wiley and Sons, Inc., New York.

Williamson, R. C., Smola, A. J., & Schölkopf, B. (2002). Entropy numbers of linear function classes. *Proc. 13th Annu. Conf. Computational Learning Theory, in press.*

Yosida, K. (1978). *Functional Analysis* (5th edition). Springer-Verlag, New York.

Appendix A: Measurability of the extension

Lemma 13. *Let $\Delta > 0$. Let X be a Banach space and $V = L_p(X)$ for some $1 \leq p < \infty$ and let $\hat{X} = X \times V$ denote the direct sum. Give both $Z = (X, Y)$ and $\hat{Z} = (\hat{X}, Y)$ the usual product metric and equip both with the Borel σ-algebra for their metric topologies. Then the map*

$$(x, y) \mapsto (x, \Delta\boldsymbol{\kappa}(x), y)$$

is measurable.

Proof. We prove the map $x \mapsto (x, \boldsymbol{\kappa}(x))$ is Borel measurable. The result then follows for the product spaces easily.

It is easy to see that any $v \in L_p(X)$ has a countable support. Consequently any point in $\hat{X}$ can be represented as $\hat{x} = (x_0, \sum_{i=1} a_i \kappa_{x_i})$ for some countable set of values $a_i, i = 1, ..,$ and some countable set of unique $x_i \in X, i = 1, ...$ Consider the preimage of the radius ϵ open ball about such an $\hat{x}$

$$\{x : |(x, \boldsymbol{\kappa}(x))) - \hat{x}| < \epsilon\}$$

which is defined by

$$|x - x_0|^2 + |\boldsymbol{\kappa}(x) - \sum_i a_i \boldsymbol{\kappa}(x_i)|^2 < \epsilon^2.$$

If x is not in $\{x_i, i = 1, ..\}$, $|\boldsymbol{\kappa}(x) - \sum_i a_i \boldsymbol{\kappa}(x_i)|^2$ is constant in x so this condition is an open or empty ball about x_0. Consequently the preimage is a (possibly empty) open ball minus at most a countable set of points and so is open and so Borel measurable. Since the open balls form a base for the metric topology and the process of taking preimages commutes with unions, the preimage of any open set is a Borel set. Let $\mathcal{E}$ denote the Borel sets whose preimage is Borel measurable and let $\mathcal{O}$ denote the sets of the topology of $\hat{X}$. Since the process of taking preimages commutes with unions, intersection, and complementation, $\mathcal{E}$ is a σ-algebra and since it contains $\mathcal{O}$ and is contained in the Borel and the Borel is the smallest containing $\mathcal{O}$ it must be the Borel. The proof is finished. ♦

Appendix B: Completion of Theorem 7

Here we complete the proof of Theorem 7 by choosing β so that $\epsilon_*(m,\delta,\beta)$ is so small that we don't mind it in $\max\left(\acute{\epsilon}(m,\delta,J^*_{z^m}),\epsilon_*(m,\delta,\beta)\right)$. Since the first term in the expression 49 for $\epsilon_*(m,\delta,\beta)$ is independent of m we choose $\beta(\mathfrak{e})$ to solve

$$\mathfrak{e} = \frac{4C^2(R^2+\Delta^2)\beta}{\mu^2}\log\left(1+\frac{\mu^2}{2C^2(R^2+\Delta^2)\beta}\right) \tag{62}$$

for small $\mathfrak{e}$ and then write

$$\begin{aligned}\epsilon_1(m,\delta,s) =& \epsilon(m,\delta,s,\beta(\mathfrak{e}))\\ =& \frac{4\alpha C^2(R^2+\Delta^2)s}{\mu^2}\log\left(1+\frac{\mu^2}{2\alpha C^2(R^2+\Delta^2)s}\right)\\ &+\frac{\log\left(\frac{4+4(\alpha m s)^{\frac{1}{2}}R}{(1-\mu)}+1\right)+\log\left(2+\frac{\log\frac{1}{\beta(\epsilon)\Delta^2}}{\log\alpha}\right)+\log\frac{1}{\delta}+2}{m}\end{aligned} \tag{63}$$

$$\begin{aligned}\epsilon_{*1}(m,\delta,\mathfrak{e}) =& \epsilon_*(m,\delta,\beta(\mathfrak{e}))\\ =& \mathfrak{e}+\frac{\log\left(\frac{4+4(m\beta(\mathfrak{e}))^{\frac{1}{2}}R}{(1-\mu)}+1\right)+\log\left(2+\frac{\log\frac{1}{\beta(\epsilon)\Delta^2}}{\log\alpha}\right)+\log\frac{1}{\delta}+2}{m}.\end{aligned} \tag{64}$$

Write equation 62 as

$$\xi = \eta\log\left(1+\frac{1}{\eta}\right)$$

where $\eta = \frac{2C^2(R^2+\Delta^2)\beta}{\mu^2}$ and $\xi = \mathfrak{e}/2$. We use the following technical lemma.

Lemma 14. *Let* $\xi = \eta\log(1+\frac{1}{\eta})$ *with* $0\le\xi\le 2^{-4}$. *Then* $\xi^2\le\eta\le\frac{1}{4}\xi$.

Proof. We bootstrap. The monotonicity of the function $\eta\log(1+\frac{1}{\eta})$ and the assumption $0\le\xi\le 1$ implies that $0\le\eta\le 1$. Therefore, $\xi=\eta\log(1+\frac{1}{\eta})\ge\eta\log(1+1)=\eta$ and so $\xi\ge\eta$. Therefore, $\eta\le 2^{-4}$ and $\xi=\eta\log(1+\frac{1}{\eta})\ge\eta\log(1+2^4)\ge 4\eta$ and the first inequality is proven. In addition we have the improved inequality

$$\eta\le\frac{1}{16\log 17}.$$

The lemma is proved if we can show that

$$\eta^{\frac{1}{2}}\log\left(1+\frac{1}{\eta}\right)\le 1$$

for all $0 \leq \eta \leq \frac{1}{16 \log 17}$. To this end we define $F(\eta) = \eta^{\frac{1}{2}} \log(1 + \frac{1}{\eta})$ and show that $F(\eta) \leq 1$ when $0 \leq \eta \leq \frac{1}{16 \log 17}$. Since $F(0) = 0$ and

$$F\left(\frac{1}{16 \log 17}\right) = \frac{\log(1 + 16 \log 17)}{\sqrt{16 \log 17}} \leq \frac{\log 81}{\sqrt{64}} \leq \frac{7}{8} \leq 1$$

it is sufficient to show that the derivative $\acute{F}(\eta) \geq 0$ for $0 \leq \eta \leq \frac{1}{16 \log 17}$. Since $\log w = \frac{\ln w}{\ln 2}$,

$$\acute{F}(\eta) = \frac{1}{\ln 2}\left(\frac{1}{2}\eta^{-\frac{1}{2}} \ln\left(1 + \frac{1}{\eta}\right) - \eta^{\frac{1}{2}} \frac{\frac{1}{\eta}}{1 + \eta}\right)$$

and it follows that $\acute{F}(\eta) \geq 0$ for $0 \leq \eta \leq \frac{1}{16 \log 17}$ if and only if

$$G(\eta) = \ln\left(1 + \frac{1}{\eta}\right) - \frac{2}{1 + \eta} \geq 0$$

for $0 \leq \eta \leq \frac{1}{16 \log 17}$. Since $G(0) = \infty$ and $G(\frac{1}{16 \log 17}) = \ln(1 + 16 \log 17) - \frac{2}{1 + \frac{1}{16 \log 17}} \geq \ln 65 - 2 \geq 0$, $G(\eta) \geq 0$ for $0 \leq \eta \leq \frac{1}{16 \log 17}$ follows if the derivative $\acute{G}(\eta) \leq 0$ for $0 \leq \eta \leq \frac{1}{16 \log 17}$. Since

$$\acute{G}(\eta) = -\frac{1}{\eta(1 + \eta)} + \frac{2}{(1 + \eta)^2}$$

it follows that $\acute{G}(\eta) \leq 0$ if and only if $\frac{1}{2} \geq \frac{\eta}{1 + \eta}$ which is true for all $0 \leq \eta \leq 1$ and the proof is finished. ♦

This inequality written in terms of β through the relations $\eta = \frac{2C^2(R^2 + \Delta^2)\beta}{\mu^2}$ and $\xi = \mathfrak{e}/2$ is

$$\frac{\mathfrak{e}^2}{8} \frac{\mu^2}{C^2(R^2 + \Delta^2)} \leq \beta(\mathfrak{e}) \leq \frac{\mathfrak{e}}{16} \frac{\mu^2}{C^2(R^2 + \Delta^2)}. \tag{65}$$

We use the inequality

$$\log\left(\frac{4 + 4(\xi)^{\frac{1}{2}} R}{(1 - \mu)} + 1\right) \leq \log\left(\frac{5 + 4(\xi)^{\frac{1}{2}} R}{(1 - \mu)}\right)$$

for any ξ, along with inequalities 65 to bound and simplify equations 63 and 64;

$$\begin{aligned} \epsilon_1(m, \delta, s) \leq & \varepsilon(m, \delta, s) \\ & = \frac{4\alpha C^2(R^2 + \Delta^2)s}{\mu^2} \log\left(1 + \frac{\mu^2}{2\alpha C^2(R^2 + \Delta^2)s}\right) \\ & \quad + \frac{\log\left(5 + 4(\alpha m s)^{\frac{1}{2}} R\right) + \log\left(2 + \frac{\log \frac{8C^2(R^2 + \Delta^2)}{\mu^2 \mathfrak{e}^2 \Delta^2}}{\log \alpha}\right) + \log \frac{1}{\delta} + \log \frac{1}{1 - \mu} + 2}{m} \end{aligned} \tag{66}$$

and

$$\epsilon_{*1}(m,\delta,\mathfrak{e}) \leq \varepsilon_*(m,\delta,\mathfrak{e}) = \mathfrak{e} + \frac{\log\left(5+\sqrt{\mathfrak{e}}\frac{\mu}{C}\sqrt{\frac{m}{(R^2+\Delta^2)}}R\right) + \log\left(2+\frac{\log\frac{8C^2(R^2+\Delta^2)}{\mu^2\mathfrak{e}^2\Delta^2}}{\log\alpha}\right) + \log\frac{1}{\delta} + \log\frac{1}{1-\mu} + 2}{m}. \tag{67}$$

Finally we note that the complement $J^*_{z^m} > \frac{\mu^2}{2\alpha C^2(R^2+\Delta^2)}$ of the event in the result 47 implies that

$$\varepsilon(m,\delta,J^*_{z^m}) \geq 1$$

because $\eta\log(1+\frac{1}{\eta}) \geq 1$ if $\eta \geq 1$. Therefore the result is also correct on the complement and the proof of Theorem 7 is finished.

Appendix C: Bounds for σ-norm and other learning strategies

Here we bound the the covering numbers of $\mathcal{L}_{r,s}$ in terms of covering numbers of operators from $X^* \to l^{2m}_\infty$ and $V^* \to l^{2m}_\infty$. We do this for a large class of criteria containing the σ-norm soft margin criteria. We proceed by covering the relevant sets by unions of products. Recall the definition of the pseudometric spaces $\mathfrak{B}(K^*)_{r,s}$ for a Banach space K defined at the beginning of Section 4.

Theorem 12. *Let X and V be Banach spaces and let $0 \leq r \leq s$ be fixed. Let $\hat{X} = X \times V$ denote the direct sum Banach space with norm $|(x,v)|^2 = |x|^2 + |v|^2$. Let $\hat{X}^* = X^* \times V^*$ denote dual space of $\hat{X}$. Consider the function*

$$J_{\mathcal{L}} : \mathcal{L} \to \Re^+$$

defined by

$$J_{\mathcal{L}}(x^*,v^*) = |x^*|^\varrho + c|v^*|^\xi$$

for some ϱ, ξ, and c. Let N be a natural number and let $(\Omega_1,\Omega_2) \in X \times V$ be a product. Let the $2m$ sample $(x,v)^{2m}$ be chosen from (Ω_1,Ω_2). Then for any $0 \leq \alpha_i \leq 1, i = 1,..,N$,

$$\mathcal{N}(\epsilon,\mathcal{L}_{r,s},2m,(\Omega_1,\Omega_2))$$

$$\leq \mathcal{N}\left(\epsilon,\mathfrak{B}(V^*)_{(\frac{r}{c})^{\frac{1}{\xi}},(\frac{s}{c})^{\frac{1}{\xi}}},2m,\Omega_2\right)$$

$$+\sum_{i=1}^{N}\mathcal{N}\left(\alpha_i\epsilon,\mathfrak{B}(X^*)_{\left((i-1)\delta\right)^{\frac{1}{\varrho}},\left(i\delta\right)^{\frac{1}{\varrho}}},2m,\Omega_1\right)\mathcal{N}\left((1-\alpha_i)\epsilon,\mathfrak{B}(V^*)_{(\frac{r-i\delta}{c})^{\frac{1}{\xi}},(\frac{s-(i-1)\delta}{c})^{\frac{1}{\xi}}},2m,\Omega_2\right).$$

Proof. Write $J_{\mathcal{L}} = \mu + \nu$ where $\mu = |x^*|^{\varrho}$ and $\nu = c|v^*|^{\xi}$. Then

$$\mathcal{L}_{r,s} = \{r < \mu + \nu \leq s\}.$$

We want to cover this set with a union of rectangles. Let $\delta = \frac{s}{N}$ and define the collection of sets

$$\{(i-1)\delta < \mu \leq i\delta\}, \quad i = 1, .., N.$$

Since $\mu \geq 0$, when $r < \mu + \nu \leq s$ it follows that $0 \leq \mu \leq s$. Consequently,

$$\{r < \mu + \nu \leq s\} = \{r < \mu + \nu \leq s\} \cap \Big(\cup_{i=1}^{N} \{(i-1)\delta < \mu \leq i\delta\} \cup \{\mu = 0, r < \nu \leq s\} \Big)$$

$$= \cup_{i=1}^{N} \Big(\{r < \mu + \nu \leq s\} \cap \{(i-1)\delta < \mu \leq i\delta\} \Big) \cup \{\mu = 0, r < \nu \leq s\}$$

However since

$$\{r < \mu + \nu \leq s\} \cap \{(i-1)\delta < \mu \leq i\delta\} \subset \{(i-1)\delta < \mu \leq i\delta\} \cap \{r - i\delta < \nu \leq s - (i-1)\delta\}$$

we obtain that

$$\{r < \mu + \nu \leq s\} \subset \cup_{i=1}^{N} \Big(\{(i-1)\delta < \mu \leq i\delta\} \cap \{r - i\delta < \nu \leq s - (i-1)\delta\} \Big) \cup \{\mu = 0, r < \nu \leq s\}.$$

Since $\mathcal{N}$ is subadditive under unions,

$$\mathcal{N}(\epsilon, \{r < J_{\mathcal{L}} \leq s\}, 2m, (\Omega_1, \Omega_2)) \leq$$

$$\mathcal{N}(\epsilon, \{\mu = 0, r < \nu \leq s\}, 2m, (\Omega_1, \Omega_2))$$

$$+ \sum_{i=1}^{N} \mathcal{N}\Big(\epsilon, \{(i-1)\delta < \mu \leq i\delta\} \cap \{r - i\delta < \nu < s - (i-1)\delta\}, 2m, (\Omega_1, \Omega_2) \Big).$$

Since the dual space $\hat{X}^*$ acts like

$$(x^*, v^*) \cdot (x, v) = x^* \cdot x + v^* \cdot v$$

and (Ω_1, Ω_2) is a product, it is clear that

$$\mathcal{N}(\epsilon, \{\mu = 0, r < \nu \leq s\}, 2m, (\Omega_1, \Omega_2)) = \mathcal{N}\Big(\epsilon, \{r < \nu \leq s\}, 2m, \Omega_2 \Big).$$

In addition, we can apply Lemma 5 to the two classes of functions $\{(i-1)\delta < \mu \leq i\delta\}$ and $\{r - i\delta < \nu < s - (i-1)\delta\}$ to obtain that for any $0 \leq \alpha_i \leq 1$

$$\mathcal{N}\Big(\epsilon, \{(i-1)\delta < \mu \leq i\delta\} \cap \{r - i\delta < \nu < s - (i-1)\delta\}, 2m, (\Omega_1, \Omega_2) \} \Big)$$

$$\leq \mathcal{N}\left(\alpha_i \epsilon, \{(i-1)\delta < \mu \leq i\delta\}, 2m, \Omega_1\right) \mathcal{N}\left((1-\alpha_i)\epsilon, \{r - i\delta < \nu < s - (i-1)\delta\}, 2m, \Omega_2\right))$$

Putting this all together in terms of the variables defined by $\mu = |x^*|^{\varrho}$ and $\nu = c|v^*|^{\xi}$ finishes the proof. ♦

We use Theorem 12 to provide bounds in terms of covering numbers of linear operators. Recall the definition (35)

$$\mathcal{N}\big(\epsilon, \Lambda \subset L(K^*, l_\infty^{2m})\big)$$

of the supremum of the covering numbers of operators constrained to $\mathcal{K} \in \Lambda$.

Theorem 13. *Let $0 \leq r \leq s$. Under the assumptions of Theorem 12 let the linear operators $\mathcal{X} : X^* \to l_\infty^{2m}$ and $\mathcal{V} : V^* \to l_\infty^{2m}$ be determined as described in Lemma 8 from a $2m$-sample $(x, v)^{2m}$.*

Then for any $0 \leq \alpha_i \leq 1, i = 1, .., N,$

$$\begin{aligned}
&\mathcal{N}(\epsilon, \mathcal{L}_{r,s}, 2m, (\Omega_1, \Omega_2)) \\
&\leq \mathcal{N}\left(\epsilon(\frac{c}{s})^{\frac{1}{\xi}}, \Omega_2^{2m} \subset L(V^*, l_\infty^{2m})\right) \\
&+ \sum_{i=1}^{N} \mathcal{N}\left(\alpha_i \epsilon (i\delta)^{-\frac{1}{\varrho}}, \Omega_1^{2m} \subset L(X^*, l_\infty^{2m})\right) \mathcal{N}\left((1-\alpha_i)\epsilon(\frac{s-(i-1)\delta}{c})^{-\frac{1}{\xi}}, \Omega_2^{2m} \subset L(V^*, l_\infty^{2m})\right).
\end{aligned}$$

Proof. To our knowledge, sharp bounds on the covering numbers of annuli do not exist. However we can bound them by the covering numbers of the closed ball

$$\mathcal{N}(\varepsilon, \mathfrak{B}(K^*)_{\rho,\sigma}, 2m, \Omega_K \subset K) \leq \mathcal{N}(\varepsilon, \mathfrak{B}(K^*)_\sigma, 2m, \Omega_K \subset K)$$

but Lemma 8 shows that

$$\mathcal{N}(\varepsilon, \mathfrak{B}(K^*)_\sigma, 2m, \Omega_K) = \mathcal{N}\left(\frac{\varepsilon}{\sigma}, \Omega_K^{2m} \subset L(K^*, l_\infty^{2m})\right).$$

Substitution of this fact for both $K = X$ and V into the result of Lemma 12 finishes the proof. ♦

We mention that since $0 \leq \alpha_i \leq 1$ and N are arbitrary, α_i can be chosen to minimize or approximately minimize the $i+1$-th term in this sum and then N can be chosen to minimize this sum. However, this is more effectively performed after specific bounds for these covering numbers are utilized.

To complete the general picture we need bounds on the covering numbers for linear operators

$$T : S \to l_\infty^{2m}$$

where S is a Banach space and we need to estimate the worst case defined in 35. However, if the domain restriction $\Omega_1 \subset B_R(X)$ for some R is used these operators are bounded by R. Likewise if there exists an R_2 such that $|\kappa_x|_V \leq R_2$ for all $x \in \Omega_2$ the operators $\mathcal{V}$ corresponding the second component are bounded by R_2. If we use bounds which are expressible in terms of the norm of the operator then the worst case bound 35 can then be estimated simply. In this case, the following corollary is very useful.

Corollary 2. *With the assumptions of Theorem* 13*, suppose that* $\Omega_1 = B_{R_1}(X)$ *and* $\Omega_2 = B_{R_2}(V)$*. Define* $L_R(K^*, l_\infty^{2m}) \subset L(K^*, l_\infty^{2m})$ *to be those linear operators with norm* $\leq R$*. Then for any* $0 \leq \alpha_i \leq 1, i = 1, .., N$*,*

$$\mathcal{N}(\epsilon, \mathcal{L}_{r,s}, 2m, (\Omega_1, \Omega_2)) \leq \mathcal{N}\Big(\tfrac{\epsilon}{R_2}(\tfrac{c}{s})^{\frac{1}{\xi}}, L_1(V^*, l_\infty^{2m}) \subset L(V^*, l_\infty^{2m})\Big)$$
$$+ \textstyle\sum_{i=1}^{N} \mathcal{N}\Big(\alpha_i \tfrac{\epsilon}{R_1}(i\delta)^{-\frac{1}{\varrho}}, L_1(X^*, l_\infty^{2m}) \subset L(X^*, l_\infty^{2m})\Big)\mathcal{N}\Big((1-\alpha_i)\tfrac{\epsilon}{R_2}(\tfrac{s-(i-1)\delta}{c})^{-\frac{1}{\xi}}, L_1(V^*, l_\infty^{2m}) \subset L(F^*, l_\infty^{2m})\Big).$$

Proof. The definitions of Ω_1 and Ω_2 imply that $|\mathcal{X}| \leq R_1$ and $|\mathcal{V}| \leq R_2$. Since $L_R(K^*, l_\infty^{2m}) = RL_1(K^*, l_\infty^{2m})$ the result follows. ♦

There are many strong results in terms of the operator norm when S is a Banach space. See for example (Carl, 1985) for general results and for strong results when S is a Banach space of type p. For the σ-norm soft margin picture we need bounds on the covering numbers of X^* and $L_\sigma(X)$. Indeed, (Carl, 1985) has strong results for l_p^N for finite N, which if extended to the $N = \infty$ case would be useful here and we have shown good bounds in Section 9.1 when X is a Hilbert space.